DOCUMENT RESUME

ED 443 297 FL 026 337

AUTHOR Hau, Kit-Tai; Marsh, Herbert W.; Kong, Chit-Kwong; Poon, Andrew Chung-Shing

TITLE Late Immersion and Language of Instruction (English vs. Chinese) in Hong Kong High Schools: Achievement Growth in Language and Nonlanguage Subjects.

PUB DATE 2000-04-00

NOTE 37p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 24-28, 2000).

PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.

DESCRIPTORS Academic Achievement; *Achievement Gains; *Chinese; *English (Second Language); Geography; High School Students; High Schools; History; *Immersion Programs; *Language of Instruction; Longitudinal Studies; Mathematics; Sciences; Second Language Instruction; Second Language Learning; Secondary Education

IDENTIFIERS Academic Language; *Content Area Teaching; *Hong Kong

ABSTRACT

This study examines how instruction in the first language (L1), Chinese, and in the second language (L2), English, affects a large sample of students' academic achievement in L1, L2, and content, nonlanguage school subjects, including mathematics, science, geography, and history, in their first 3 years of high school. For all four content area subjects, to a lesser extent in mathematics, late immersion in English as the language of instruction had negative effects that did not vary with initial general ability; were slightly smaller for students initially more proficient in the second language; declined slightly over time for some subjects; and were counteracted somewhat by particularly strong English-language courses. Immersion in English had positive effects on English and, to a lesser extent, Chinese language achievement, but these effects were small relative to the large negative effects in nonlanguage subjects. Whereas previous research has shown positive effects for early-immersion programs that start in kindergarten where language demands are not so great, negative effects for this late immersion program challenge the generality of these findings to high schools, and perhaps even theoretical models of second language acquisition. The paper begins with a literature review, provides a description of the situation in Hong Kong, presents empirical data, and concludes with an exploration of the educational policy implications. (Contains 45 references.) (KFT)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 443 297



# Late Immersion and Language of Instruction (English vs. Chinese) in Hong Kong High Schools: Achievement Growth in Language and Nonlanguage Subjects

Kit-Tai Hau, The Chinese University of Hong Kong, Hong Kong
Herbert W. Marsh, University of Western Sydney, Macarthur, Australia
Chit-Kwong Kong, The Chinese University of Hong Kong, Hong Kong
Andrew Chung-Shing Poon, Hong Kong Education Department

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Prof. Kit Tai Hau

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.



FL026337

BEST COPY AVAILABLE

# Late Immersion and Language of Instruction (English vs. Chinese) in Hong Kong High Schools: Achievement Growth in Language and Nonlanguage Subjects

Kit-Tai Hau, The Chinese University of Hong Kong, Hong Kong
Herbert W. Marsh, University of Western Sydney, Macarthur, Australia
Chit-Kwong Kong, The Chinese University of Hong Kong, Hong Kong
Andrew Chung-Shing Poon, Hong Kong Education Department

Paper presented at American Educational Research Association Annual Meeting, April 24-28, New Orleans.

**Abstract** In this article, the effects of instruction in the first language (Chinese) and the second language (English) on achievement were evaluated using multilevel growth models for a large representative sample of Hong Kong students during their first three years of high school. For nonlanguage subjects (history, geography, science, and, to a lesser extent, mathematics), late immersion in English as the language of instruction had large negative effects that: (a) did not vary with initial general ability, (b) were slightly smaller for students initially more proficient in the second language, (c) declined slightly over time for some subjects, and (d) were counteracted somewhat by particularly strong English-language classes. Immersion in English had positive effects on English and, to a smaller extent, Chinese language achievement, but these effects were small relative to the large negative effects in nonlanguage subjects. Whereas previous research has shown positive effects for early-immersion programs that start in kindergarten where language demands are not so great, negative effects for this late-immersion program challenge the generality of these findings to high schools and, perhaps, theoretical models of second-language acquisition.

This study examines how instruction in the first language (L1, Chinese) and in the second language (L2, English) affects high school students' achievement in L1, L2, and four content (nonlanguage) school subjects (mathematics, science, geography, and history). This research has important policy implications in Hong Kong, but it also has theoretically and empirically important implications for understanding immersion strategies in other contexts, especially where students are instructed in L2 for all (or most) school subjects. Although early-immersion programs that begin at the start of formal schooling have a long history of success (e.g., Lambert, 1990, 1992), there has been much less systematic, large-scale research on late-immersion programs that begin in high school, particularly in relation to achievement in nonlanguage school subjects. We begin by reviewing previous empirical and theoretical research that informs our study. Next, we describe the Hong Kong context in which this research was conducted, present results of this large-scale, longitudinal study on second-language instruction, and, finally, explore policy implications for the Hong Kong educational system and for immersion programs more generally.

## Empirical and Theoretical Background

In the twentieth century, research on second-language acquisition and instruction throughout the world has been largely driven by the applied needs of education policy and practice and has been informed by disciplines as diverse as education, linguistics, sociolinguistics, psycholinguistics, cognitive science, psychology, educational psychology, crosscultural psychology, anthropology, popular culture, sociology, and others (e.g., Cummins, 1979, 1986, 1991, 1996; Francis, 1999; Garcia, 1993; Hakuta, 1986; Hakuta & McLaughlin, 1996; Lambert, 1992; Swain & Johnson, 1997; Willig, 1985). Two movements are particularly relevant to our study: the transitional bilingual programs and the Canadian immersion programs.

ED 426337

BEST COPY AVAILABLE

# Late Immersion and Language of Instruction in Hong Kong High Schools: Achievement Growth in Language and Nonlanguage Subjects

Herbert W. Marsh
University of Western Sydney, Macarthur, Australia
Kit-Tai Hau & Chit-Kwong Kong
The Chinese University of Hong Kong, Hong Kong

In this article, Herbert Marsh, Kit-Tai Hau, and Chit-Kwong Kong evaluate the effects of instruction in the first language (Chinese) and the second language (English) on achievement using multilevel growth models for a large representative sample of Hong Kong students during their first three years of high school. For nonlanguage subjects (history, geography, science, and, to a lesser extent, mathematics), late immersion in English as the language of instruction had large negative effects that: (a) did not vary with initial general ability, (b) were slightly smaller for students initially more proficient in the second language, (c) declined slightly over time for some subjects, and (d) were counteracted somewhat by particularly strong English-language classes. Immersion in English had positive effects on English and, to a smaller extent, Chinese language achievement, but these effects were small relative to the large negative effects in nonlanguage subjects. Whereas previous research has shown positive effects for early-immersion programs that start in kindergarten where language demands are not so great, negative effects for this late-immersion program challenge the generality of these findings to high schools and, perhaps, theoretical models of second-language acquisition.

This study examines how instruction in the first language (L1, Chinese) and in the second language (L2, English) affects high school students' achievement in L1, L2, and four content (nonlanguage) school subjects (mathematics, science, geography, and history). This research has important policy implications in Hong Kong, but it also has theoretically and empirically important implications for understanding immersion strategies in other contexts, especially where students are instructed in L2 for all (or most) school subjects. Although early-immersion programs that begin at the start of formal schooling have a long history of success (e.g., Lambert, 1990, 1992), there has been much less systematic, large-scale research on late-immersion programs that begin in high school, particularly in relation to achievement in nonlanguage school subjects. We begin by reviewing previous empirical and theoretical research that informs our study. Next, we describe the Hong Kong context in which this research was conducted, present results of this large-scale, longitudinal study on second-language instruction, and, finally, explore policy implications for the Hong Kong educational system and for immersion programs more generally.

## Empirical and Theoretical Background

In the twentieth century, research on second-language acquisition and instruction throughout the world has been largely driven by the applied needs of education policy and practice and has been informed by disciplines as diverse as education, linguistics, sociolinguistics, psycholinguistics, cognitive science, psychology, educational psychology, crosscultural psychology, anthropology, popular culture, sociology, and others (e.g., Cummins, 1979, 1986, 1991, 1996; Francis, 1999; Garcia, 1993; Hakuta, 1986; Hakuta & McLaughlin, 1996; Lambert, 1992; Swain & Johnson, 1997; Willig, 1985). Two movements are particularly relevant to our study: the transitional bilingual programs and the Canadian immersion programs.

*Transitional Bilingual Programs: Additive Bilingualism vs. Submersion*

Bilingual and second-language studies are often politically controversial, policy-oriented, applied research that is directed at real-world problems. Beginning in the 1980s, particularly in the United States, this research was fueled by political and educational needs arising from large immigrations. It focused on whether minority native languages should be "submerged" (i.e., replaced with the dominant English language), or whether students with limited English proficiency should receive English-language instruction and instruction in other subjects in their own native language until they achieve competence in English (i.e., transitional bilingual programs; see Willig, 1985). Submersion programs have typically focused on the extent and speed of assimilation and the acquisition of the new dominant language; they have typically placed little emphasis on maintenance of the native language. Following the controversial 1974 U.S. Supreme Court *Lau v. Nichols* decision, the critical "Lau question" became whether there was sufficient research evidence to mandate transitional bilingual programs in which non-English-speaking or limited-English-proficient students are taught in their native language until they reach an appropriate level of proficiency in English.[1]

A decade later, Willig (1985) conducted a sophisticated meta-analysis, comparing transitional bilingual education programs in the United States with traditional programs (typically, "submersion" programs in which non-native speakers are taught exclusively in English). In order to juxtapose traditional narrative and meta-analytic reviews, Willig focused on studies from the Baker and de Kanter (1981) review of research, but limited her consideration to U.S. studies. Controlling for prior student differences and methodological inadequacies, students in bilingual programs who were taught in their first language performed better than students taught in L2 on language and nonlanguage achievement tests in both English and L1, and had better attitudes toward self and school. The differences favoring transitional bilingual programs over submersion programs were systematically larger for studies that Willig judged to be methodologically stronger, based on traditional criteria used in meta-analyses. The differences were, however, smaller in methodologically weaker studies in which comparison students were not matched to bilingual students (e.g., non- or limited-English-proficient-students in bilingual programs vs. English-proficient students in comparison programs). Whereas L1 (non-English) language test scores were substantially higher in the transitional bilingual programs, scores in other school subjects were also higher, as were, to a lesser extent, L2 (English) skills. Although the effects were positive for all language components, they were smaller for oral language than for writing, listening, reading, and vocabulary.

Most research in Willig's 1985 review emphasized language achievement, but the largest effects were for social studies achievement. This finding led Willig to suggest that "bilingual education may be succeeding in preventing academic lag in language-mediated academic subjects, but, unfortunately, these [subjects] are seldom included in evaluation designs" (p. 311). In a similar vein, she emphasized that science comprehension is linked to understanding the language in which it is presented, but lamented that not a single study in her meta-analysis included science achievement. Hence, the failure to evaluate the effects of language of instruction on achievement in both nonlanguage content and language subjects is a critical limitation in this past research.

More recent research (e.g., Collier, 1992; Cummins, 1996; Greene, 1998; Krashen, 1997) provides further support for well-designed transitional bilingual programs. Krashen (1997) argued that good

---

[1] The *Lau* decision was based on a claim that the needs of Chinese Americans with limited English proficiency had not been met under the Civil Rights Act of 1964. The ruling would have mandated the implementation of transitional bilingual education programs for students with limited English proficiency, but it was subsequently withdrawn. For further discussion, see Hakuta and McLaughlin, 1996.



BEST COPY AVAILABLE

bilingual programs provide students with content knowledge and literacy in L1, which indirectly aids L2 proficiency, because it is easier to learn to read in the L1 language that a child already knows. He concluded that the best bilingual programs provide initial L1 instruction, L2 language classes, sheltered classes in which students with intermediate L2 skills are taught nonlanguage subjects in L2, and a gradual transition to mainstream nonlanguage content classes taught in L2. Focusing on longitudinal studies, Collier (1992) argued that student achievement increases over time with the amount of L1 instruction language-minority students receive in combination with L2 support when compared to matched groups taught entirely in L2. While recognizing that bilingual education per se is no panacea, Cummins (1996) concluded that "both large-scale and small-scale studies consistently show that strong promotion of bilingual students' L1 throughout elementary school contributes significantly to academic success" (p. 121).

## Canadian Immersion Programs: Additive Bilingualism

### Early history of the Canadian Immersion Program

Lambert (1990, 1992) summarized his research about what has become known as the Canadian immersion program. In Quebec, in the 1950s, French Canadians and English Canadians were distinct groups, separated by social status as well as language. Even though French Canadians were a majority in Quebec, they had lower status jobs than English Canadians and used English to some extent, whereas English Canadians had little need or desire to learn French. Lambert and his colleagues reasoned that learning a foreign language depended not only on general intellect and language aptitude, but also on positive perceptions of the target language group (e.g., the French Canadians in their research). They developed a French-language immersion program specifically devised for English Canadian children with little or no exposure to French. In this program, these children spent the first three years of schooling (beginning in kindergarten) learning almost exclusively in French. English was first introduced in second or third grade and gradually increased to about 50 percent of instruction time. The aim of the Canadian immersion program was to achieve additive bilingualism – the acquisition of a new L2 while maintaining or enhancing L1 proficiency and achievement in other school subjects. Based on consistent findings from over twenty-five years, Lambert (1992; also see 1981, 1990) argued that immersion students achieve a remarkably high level of functional bilingualism and biculturalism (a cultural appreciation of French Canadians; Genesee, 1984). Lambert also contended that bilinguals develop divergent thinking skills. In addition, bilinguals acquire English-language skills by their upper primary school years (even though they have spent much less learning time in English; see Cummins, 1979) and are able to achieve in some content areas at levels comparable to that of non-immersion students. Growing evidence further suggested that the advantages produced by the immersion program generalized for English Canadian students differing in socioeconomic status and IQ (see Genesee, 1987).

Genesee (1985) argued that immersion is not so much a method of L2 acquisition, but a pedagogical approach that promotes L2 acquisition in which L2 instruction in nonlanguage content subjects creates conditions like those in L1 instructional contexts. Genesee also emphasized that immersion programs can be distinguished by two factors of time: the initiation (early, kindergarten; delayed, grade 4 or 5; late, grade 7 or 8) and the extent to which instruction is provided in both L1 and L2 (total, in which all or most instruction is in L2, or partial [mixed]). He then summarized preliminary or formative evaluations of several variations to the traditional immersion program, suggesting that accumulated research evidence demonstrated that gains in L2 proficiency were not at the expense of L1 proficiency. He emphasized, however, that "immersion programs were designed for English-speaking, majority group children, and the evaluation results pertain to this population only" (Genesee, 1985, p. 556). Genesee (1978) also reviewed empirical and theoretical issues having to do with the optimal starting time for immersion programs and

concluded that, whereas benefits can result from late-immersion programs, "high levels of second language proficiency are best achieved by an early start and long duration of instruction, provided that effective teaching methods are employed" (p. 1). There are, however, unresolved issues in studies of immersion programs that we now consider.

*The Prototypical Immersion Program and Unresolved Issues*

Swain and Johnson (1997) focused specifically on how immersion programs differ from other types of bilingual and L2 education programs. This issue is important to our claim that our research is an immersion study. They offered the following defining characteristics of an immersion study: (a) L2 is the medium of instruction, (b) the immersion curriculum parallels the local first-language curriculum; (c) overt support exists for L1; (d) the program aims for additive bilingualism; (e) L2 exposure is largely confined to the classroom; (f) students enter with similar and limited levels of L2 proficiency; (g) teachers are bilingual; and (h) the classroom culture is that of the local L1 community. Within this context, Johnson (1997) argued that instruction in English in Hong Kong high schools, the focus of the present investigation, constitutes a late-immersion program.

Swain and Johnson (1997) identified a variety of ongoing, unresolved concerns. For example, a practical problem common to most immersion programs is how to attract and retain the highly committed bilingual teachers who have appropriate expertise in both the language of instruction and the subject content area. The selection of students for participation in immersion programs is also a controversial issue. Swain and Johnson question whether the selection of immersion students is justifiable or an unnecessary form of discrimination. They emphasize that "it seems likely that L1 literacy, general academic achievement, L2 proficiency, and motivation might all become increasingly important the later an immersion program begins" (p. 13). They also recognized that at higher levels of education and in more abstract content areas, L2 instruction might need to be supplemented with L1 instruction, but they also indicated that further research was needed into possible strategies to best meet this apparent need.

Swain and Johnson (1997) are surprisingly silent on the critical issue of the effects of immersion on achievement in content subjects (other than language), particularly in late-immersion programs. The language limitations in what can be presented to students (comprehensible input) and students' inability to grapple with complex and abstract ideas in a L2 they have not yet mastered can place students at a distinct disadvantage relative to what they would have achieved if taught in their L1. Although some researchers claim that immersion students achieve at levels comparable to non-immersion students, such comparisons are often based on performances of an initially elite group of immersion students relative to representative norms and do not control for initial ability differences (e.g., Duff, 1997). Researchers (e.g., Duff, 1997; Met & Lorenz, 1997) have noted that, whereas immersion students may be able to master the concrete objectives required in the primary (K-2) curriculum, their ability to explore more abstract concepts in nonlanguage subjects as required in subsequent levels of education is frustrated by limitations in L2 proficiency.

Swain and Johnson (1997) concluded that under favorable conditions, immersion results in additive bilingualism and provides cognitive, cultural, and psychological advantages as well. They also noted, however, that under less favorable conditions, immersion programs may not be able to achieve their full potential benefits. Although it might be possible merely to dismiss any negative results as not representing true immersion research, Swain and Johnson recognized that a better understanding of why such programs have not resulted in the expected benefits is needed.

*Theoretical Models of Bilingual Proficiency*

BEST COPY AVAILABLE

Why are outcomes typically more positive in immersion-type programs like those summarized by Lambert (1992) than in submersion-type programs like those summarized by Willig (1985)? Because of the strong pragmatic and policy orientations of this research, many approaches to this question have not been based on a well-grounded theoretical perspective that is able to generate testable hypotheses. Several models of bilingual proficiency (see reviews by Bourhis, 1990; Gardner, 1985) have had considerable impact. However, for this study, we focus on two models (Cummins, 1979; Lambert, 1974) that are most relevant to our research.

*Lambert s (1974) Social Psychology Model of Second-Language Learning*

Bourhis (1990; see also Gardner, 1985) argued that Lambert's model is the precursor of the social process models that focus on the social factors that motivate or demotivate students to learn a second language. In Lambert's model, second-language learning is a function of (a) prior aptitude (including general cognitive ability and specific language skills) and (b) motivation, which is itself a function of prior attitudes about the language and the people who speak the language and motivational orientations for learning the language. A critical feature of Lambert's model is his emphasis on self-identity. According to Lambert, once bilingualism has developed sufficiently, it will influence self-identity in a manner dependent upon whether or not students feel that they belong to a desired group of people. To the extent that L2 proficiency is not intended to substitute for L1 proficiency (as in the Canadian immersion programs), additive bilingualism is predicted to occur and the effects on self-identity are predicted to be positive. However, when the L2 is intended to replace L1, subtractive bilingualism is predicted to occur that may lead to social alienation.

*Cummins' Theory of the Interaction Between Student Characteristics and Program Type*

Cummins presents an aptitude-treatment-interaction theory in which the effectiveness of the intervention (the type of program) depended on characteristics of the students participating in the program (aptitudes). More specifically, Cummins argues that "a cognitively and academically beneficial form of bilingualism can be achieved only on the basis of adequately developed first language (L1) skills" (1979, p. 222) and posits his interaction model that is based on two theoretical hypotheses (Cummins, 1979, 1986, 1991, 1996; Cummins, Harley, Swain, & Allen, 1990). The *developmental interdependence hypothesis* posits that second-language competence is partially a function of actualized L1 competence at the start of second-language instruction. Cummins and Swain (1986) more fully articulate this hypothesis, stating that "to the extent that instruction in Lx is effective in promoting proficiency in Lx, transfer of this proficiency to Ly will occur provided there is adequate exposure to Ly (either in school or environment) and adequate motivation to learn Ly" (p. 87). The *threshold hypothesis* posits that students must achieve minimum thresholds of proficiency in both languages before the benefits of bilingualism can be achieved (and to avoid detrimental effects of "semilingualism," or inadequate skills in both languages). Based on these hypotheses, Cummins predicted aptitude-treatment interactions in which outcomes are a function of the interaction between sociocultural background, child input (e.g., language competence and motivation to learn L2 and maintain L1), and educational program factors. Based on this interdependence principal, Cummins (1996) argues against a simple "time-on-task" hypothesis (that more time spent in instruction in English automatically translates into better English proficiency), claiming that development of good L1 proficiency will transfer to L2 proficiency.

Threshold considerations led Cummins (1979) to ask what levels of L2 proficiency are needed at different grade levels for students to benefit from instruction in L2, and whether continued development of L1 skills is important in the development of L2 skills. He proposed that the effects of semilingualism (subthreshold skills in both languages) would be detrimental, whereas the effects of dominant

BEST COPY AVAILABLE

bilingualism (subthreshold skills in one language but high levels in the other) might not be negative. The effects of additive bilingualism (high levels in both languages)) would be positive. He also indicated, however, that the required threshold level would vary with stage of schooling. Thus, for example, one of the reasons why no cognitive disadvantages seem to be associated with early-immersion programs is that cognitive development in early school years is less based on formal language than in later school years. In bilingual programs for limited-English-proficient students, maintenance of L1 has benefits compared to students taught entirely in English, who may never achieve competence in either language.

The developmental interdependence hypothesis predicts that L2 acquisition will be better without loss of L1 proficiency when children already have good L1 mastery, as in the Canadian immersion program. Consistent with this hypothesis, Cummins (1979) reports that home and social experiences are sufficient to attain functional L1 proficiency for most middle-class language-majority children, and that the ability to extract language from text is easily transferred from one language to another. Consistent with this hypothesis, he reviewed Swedish research showing that in Swedish schools Finnish students who immigrated to Sweden at age ten quickly achieved competence in Swedish and surpassed migrant children born in Sweden, whereas Finnish students who immigrated to Sweden at age seven or eight had serious problems. The interaction between these two hypotheses explains why native-language instruction is important in minority-language situations (where it is more difficult to maintain and develop L1), like the bilingual programs reviewed by Willig (1985), but not when L1 is the dominant language (supported in the home and community) as in the Canadian immersion programs. In addition to linguistic and educational program factors, Cummins (1979, 1986; see also Lambert, 1992) argues that motivational factors are important, as students who positively identify with both target language groups are more likely to achieve additive bilingualism, whereas those who identify with neither are more likely to suffer semilingualism.

Critical and Optimal Ages for Learning a Second Language

Second language acquisition, as emphasized in theoretical and empirical research already reviewed, is seen as building substantially on L1 skills already acquired rather than being in conflict with competing habits based on the first language (e.g., Cummins, 1979, Hakuta & McLaughlin, 1996; Lambert, 1992). From this perspective, it is important to ask whether there is an optimal age for L2 learning and whether early- or late-immersion programs are more effective. Singleton (1992; see also Hakuta & McLaughlin, 1996) argues that no clear-cut evidence exists for an optimal age for L2 learning. According to Singleton, on the one hand, students who begin formal instruction in L2 at a later age tend to "catch up" with students who begin at an earlier age. Conversely, for students learning in a naturalistic context, those with an earlier initial exposure gain greater fluency than those with a later initial exposure, even after controlling for overall exposure. Singleton, however, claims that there were other reasons for preferring early-immersion because of the "crowdedness" of the curriculum in later school years. Building on work by Krashen (1981), Singleton also argues that successful L2 learning requires comprehensible input that actively engages the learner's attention.

Cummins, Harley, Swain, and Allen (1990) and d'Anglejan (1990) reviewed research from the Development of Bilingual Proficiency project that focused on age differences and the optimal age for immersion. Based on a large study of Japanese immigrants to Canada, d'Anglejan argued that more information was needed about the ability of Japanese high school students to handle the particularly substantial reading demands in nonlanguage content subjects. She hypothesized that discrepancies in overall school performance between native and non-native speakers of English would be particularly apparent at high school levels. D'Anglejan also offered a particularly candid evaluation of Canadian

immersion research comparing the long-term results of early-immersion, late-immersion, and extended French (non-immersion) programs. According to d'Anglejan:

> [T]he discovery a few years ago that the early immersion children's head start in French did not seem to result in the systematic advantages that might have been predicted came as a surprise to many of us.....The present study confirms once again the lack of any systematic advantages ascribable to an early start. Indeed, it suggests that some good and some bad things may result from all three types of program. (pp. 152-153)

## The Present Investigation: Predictions in the Hong Kong Context

In this section, we briefly summarize the Hong Kong context and how it relates to research into immersion, bilingualism, and second-language acquisition. We then develop research questions and theoretical predictions that are pursued in the present investigation.

In Hong Kong, both Chinese and English are highly valued and important school subjects. Whereas Chinese (Cantonese) is the language of everyday use, English is used mainly for education, government, and business purposes and not usually for social discourse. The Education Commission (1990), the highest government advisory committee on all major educational policy, was established by the Hong Kong government to examine issues related to language of instruction. This commission emphasized that "there is pressure for children to learn English, since this is seen by parents as offering the best prospect for their children's future. Many children, however, have difficulty with learning in English; and conversely, Chinese is undervalued as a medium of instruction and the importance of Chinese language skills is not sufficiently recognized" (p. 93). Recognizing competing needs, the commission stressed that there was a need for some English-language high schools (i.e., schools that use English as the medium of instruction) in order to maintain Hong Kong's international position as a business, financial, and trading center. On the other hand, "since research has shown that students can study effectively in English only when they have passed a certain threshold of language competence in both their mother tongue and in English, the Working Group proposed that English-medium secondary education should be open only to those who had reached this threshold" (Education Commission, 1990, p. 94), a value suggested to be the top 30 percent of students. The report went on to emphasize that research has shown that "the majority of students will learn more effectively through their mother tongue than through English" (p. 95). From this perspective, policies were pursued to ensure that "each student was educated through a medium likely to lead to maximum cognitive and academic development. English should only be used as a medium of instruction where students could benefit from this" (p. 96). These 1990 policy recommendations were subsequently endorsed in the 1995 Educational Commission Report No. 6, which formally recommended "embarking on a comprehensive research programme to follow the academic and personal development of groups of students, matched for academic ability and experiencing different medium of instruction models" (p. xvi). On the basis of this recommendation, the present investigation was initiated. This inquiry is a large-scale, quasi-experimental research study following from these recommendations that focuses specifically on student performance in schools that use different languages of instruction.

In contrast to policy recommendations reinforcing Chinese medium of instruction, Hong Kong parents believe that English-medium instruction is most advantageous and that Chinese-medium instruction is potentially disadvantageous (see also Gibbons, 1989). Hence, many of the most prestigious and highly selective schools in Hong Kong use English as the language of instruction. Furthermore, because of parental beliefs, English-language high schools are reluctant to lose any competitive advantage by

switching to Chinese, and Chinese-language schools may experience pressure to switch to instruction in English.

Earlier work by Gibbons (1989) demonstrates that these current issues in Hong Kong are not new. Gibbons noted, for example, that initial recommendations by government to provide more instruction in Chinese in the 1960s and again in the 1970s were compromised in the face of strong parental opposition and the commercial value placed on English-language skills. Recognizing limitations in existing research, he nevertheless concluded that the cumulative evidence indicated that instruction in Chinese was more effective than instruction in English because students understood Chinese better and because instruction in English was particularly disadvantageous for lower ability students. Nevertheless, parental pressure forced some schools to become English-language schools even though the practical realities of the teaching situation forced teachers to use a mixture of English and Chinese. According to Gibbons (1989), there was general agreement among senior administrators in the Hong Kong Department of Education and university academics that more instruction should be in Chinese, but that this was not a politically viable option. This complex interplay between public policy and politics created the paradoxical situation in which a British colonial government pressed for greater emphasis on Chinese instruction but faced strong resistance from the local Chinese community which wanted more emphasis on instruction in English.

Johnson (1997) discussed specifically the Hong Kong context in relation to immersion programs at approximately the time that the present investigation was initiated. Emphasizing that many students and, perhaps, some teachers were not adequately equipped to deal with a total-immersion into instruction in English, Johnson (1997) argued that much so-called instruction in English actually was based on a "mixing and switching" mode of instruction that is a mix of English and Cantonese language use (see also Gibbons, 1989). Hence, even though many high schools claim to teach in English, there continues to be large variation in the extent to which English is actually used. Although English-language textbooks were designed to meet the requirements of a prescribed syllabus and public examinations, teachers in early high school years tended to simplify the vocabulary and discourse, emphasizing statements of fact and relying on pictures and graphs to convey meaning. Johnson (1997) reviewed Hong Kong studies evaluating the success of the immersion studies and determined that they were inconclusive. Although Johnson reported some Department of Education research showing that students taught in English and in Chinese did not differ in terms of achievement in content subjects, whereas instruction in English produced better English achievement, it is not clear that the research appropriately controlled for large initial differences in prior academic achievement. Although Johnson noted planned policies to reduce the number of high schools allowed to teach in English or to increase the entrance requirements to enroll in such schools, he also noted that these proposals have been criticized as being "elitist and socially divisive and as relegating Chinese-medium instruction to second-class status" (p. 185).

Johnson (1997) also concluded that immersion education in Hong Kong high schools "fails to produce the high level of second language proficiency that is expected from it and this is the only justification for such a programme, given that the L1 alternative is available and is in all aspects appropriate" (p. 185). He continues by noting, however, that the limited evidence available suggests that no decline in outcomes has occurred in nonlanguage content subjects relative to other developed educational systems and that good Chinese language skills have been maintained. Given that Johnson cited some research by Gibbons, it is surprising that he did not review Gibbons's earlier conclusion that instruction in English was detrimental to achievement in nonlanguage subjects. Although Johnson seems to reach different conclusions from Gibbons (1989) about the effect of instruction in English on achievement in nonlanguage classes, both agree that there is not adequate research upon which to base firm conclusions. However, Johnson also emphasizes that:

BEST COPY AVAILABLE

> whether it is better, as opposed to easier, to educate students through their native language arguably becomes irrelevant if it is a major requirement of the society the education system serves that it should produce at least some students with high levels of bilingual proficiency that can only be achieved through immersion. (1997, p. 182)

Juxtaposition Between the Hong Kong Setting and Previous Research

Although English is taught in Hong Kong primary schools (grades 1–6), the level of English competence for most students is not high (see Gibbons, 1989). At the end of primary school, students are assigned a global achievement score based on their school achievement across academic subjects. All students also take common standardized verbal (Chinese) and mathematics aptitude tests so that the ranking of students from each school can be combined to generate a common basis for ranking all grade six students (i.e., the school-based achievement is moderated by the standardized tests). In Hong Kong, students can apply to any of a wide variety of high schools where the language of instruction is mainly Chinese (except for English classes), mixed Chinese and English, or mainly English (except for Chinese classes). Using the classification scheme proposed by Genesee (1985), high schools that provide instruction in English represent partial or total late-immersion programs. Using the Swain and Johnson (1997) definition of immersion programs, Johnson (1997) notes that Hong Kong English-language high schools clearly constitute a late-immersion program. In terms of the components of the Cummins model (1979), Chinese students attending high schools taught in English typically have very good L1 (Chinese) proficiency, are highly motivated to learn L2 (English), and are also highly motivated to maintain and develop their L1 skills, which are reinforced in family and community settings. Hence, the Cummins model predicts that these students should be ideally suited for a total-immersion program and that their cognitive and academic performances should be superior to Chinese students attending Chinese-language (non-immersion) high schools, after controlling for preexisting differences.

A critical, unresolved theoretical issue is whether immersion effects depend on prior academic and language proficiency. Traditional practice in immersion programs includes explicit selection based on student characteristics such as prior achievement levels or implicit selection on the basis of parental (or student) choice. There is, however, a philosophical orientation toward more inclusive selection strategies supported by limited research, suggesting that the benefits of an immersion program are broadly generalizable (e.g., Lambert, 1992; Genesee, 1985). Theoretical models – particularly the Cummins (1979) interaction model – predict that the additive bilingualism needed in order for students to benefit from an immersion program is more likely when students have appropriate competencies and motivations to participate. Clearly, the Hong Kong Education Commission (1990) accepted the logic of this perspective. They argued that the effects of instruction in English would vary according to initial ability levels and that students who were not particularly able might be disadvantaged by being taught nonlanguage subjects in English. Following from these predictions, we specifically test the hypothesis that the effects of language of instruction vary according to prior student abilities.

Because Hong Kong implemented late immersion rather than early immersion, the present investigation makes an ideal setting for testing the generalizability of findings based on previous research and theoretical predictions from Cummins's 1979 model. Although the predictions are apparently straightforward – even more so, perhaps, than in the original Canadian immersion programs – there are some crucial differences that make the present investigation a particularly important test of the theory. Whereas the predicted advantages of English-language high schools are unproblematic for English-language achievement and, perhaps, even Chinese-language achievement, predictions are more complicated for nonlanguage subjects. As noted earlier, several authors have alluded to potential

problems of content mastery in nonlanguage high school subjects, where it may be more difficult to teach complex and abstract concepts in an L2 that has not yet been adequately mastered. Willig (1985) also noted that some of the largest advantages of teaching students in their L1 were observed in the nonlanguage areas, particularly social studies. Although no studies in her review focused on science achievement, Willig also hypothesized that students taught science in their L1 would also be advantaged over students taught science in L2. Also, whereas we interpret Cummins (1979) as predicting that students in English-language schools should also excel in all school subjects, we recognize that he qualified his predictions with the caveat that the level of L2 proficiency needed to achieve additive bilingualism (i.e., his "threshold level") increases at higher levels of schooling.

Specific Research Questions to be Addressed in the Present Investigation

As emphasized earlier, the overarching research question to be addressed is: What are the effects of language of instruction on achievement during the first three years of high school after controlling for initial differences in student achievement? Based on our review of the theoretical and research literature, and policy issues from the Hong Kong context, we now pose a series of research questions to guide our analyses and presentation of the results:

1. Do the effects of instruction in English vary substantially for different school subjects? Whereas most research has focused on the effects of instruction in L2 on the development of language skills in L1 and L2, it is also important to evaluate the effects of instruction in L2 on nonlanguage content subjects.

2. Do the effects of the language of instruction vary with prior student characteristics, such as prior achievement or prior English skills? For example, are students who initially are brighter in general or have better English skills more advantaged – or less disadvantaged – by being taught in English?

3. Do the effects of instruction in English grow larger or smaller over the first three years of high school? It might be expected, for example, that any negative effects of instruction in English might be larger in the first year of high school when students are first introduced to instruction in English, but that any negative effects might become smaller over time as students became more accustomed to instruction in English and improved their English-language skills.

4. What are the effects of English in English classes and English ethos in nonclassroom activities? In English-language high schools, students are exposed to English in all of their subjects other than Chinese. These students, however, also learn English in English-language classes and are exposed to English in nonclassroom activities (e.g., extracurricular activities, school meetings, and school notice boards). What are the effects of these other sources of exposure to English on achievement in different school subjects and how do these vary with the language of instruction?

Methods

*Sample*

In Hong Kong, the highly competitive selection into different high schools at the end of grade six is based on parental choice and on examination results.[2] Schools that attract better students are those with better public examination results, higher admission rates to universities, a longer history of positive results, a

[2] **A student's examination score (a** "secondary school places allocation" score) is placed into five broad bands of equal size (20% of the students per band). Students in the higher achievement bands are allocated by their parental choice **first. However,** students within the same band **who** apply to the same oversubscribed school are allocated randomly, so that even the most selective schools will have an achievement mix of students within the top 20 **percent** of all students. In addition, a small number of primary and secondary schools are linked so that some secondary school places are reserved for the linked primary schools if these students can meet minimum placement score standards. Within these constraints, students are free to choose any high school in Hong Kong.

BEST COPY AVAILABLE

good reputation with parents, and other desirable characteristics (e.g., school culture, extracurricular activities, proximity to home). As emphasized earlier, language of instruction is one important consideration in the selection of schools by parents, and schools with instruction in English are highly prestigious. Schools in this study had considerable freedom in choosing the language of instruction. Some schools taught all classes (other than Chinese) primarily in English, some taught all classes (other than English) in Chinese, and some used both Chinese and English. The present study is a large-scale investigation on the effects of language of instruction for secondary schools in Hong Kong. The sample is broadly representative of Hong Kong schools. The schools were selected by the Hong Kong Department of Education, using public documents and their academic subject inspectors' and other officers' knowledge of the schools to provide a large sample of schools that was representative in relation to students' academic ability and the language of instruction used by the school. This "local knowledge" about language of instruction used by the schools was subsequently validated with a survey completed by students about the language of instruction that was actually used in the school. The original sample consisted of 12,784 Chinese secondary students in grade seven attending one of fifty-six high schools. The schools were selected by the Hong Kong Department of Education to include a diverse sample of schools broadly representative of Hong Kong secondary schools in terms of religious background, mode of government subsidy, gender grouping, and, of particular relevance to the present investigation, language of instruction. For the selected schools, all students entering grade seven were included in the study.

*Procedures and Measures*

In Hong Kong, all grade six students are allocated a placement score that represents an internal aggregate of achievement in all school subjects except physical education (although Chinese, English, and mathematics are weighted more heavily) that is moderated by (i.e., adjusted in relation to performance on) external examinations. The external examinations are standardized measures of general ability, with separate mathematics and verbal (Chinese) components, that are administered by the Department of Education. Because these scores are the primary basis for the extremely important selection into high schools, performances on these achievement tests are very important to students and schools. For purposes of the present investigation, prior achievement (performance in grade six at time 0, the year prior to the start of high school) is based on five separate scores: 1) the original placement achievement score (Ach0) (i.e., the placement score before it was divided into the five categories representing the "bands" used to allocate students into high schools), 2) the mathematics moderator examination score, 3) the verbal (Chinese) moderator examination score , 4) school-based performance in Chinese, and 5) school-based performance in English.

In each of the three years following entry into high school, the Hong Kong Department of Education administered standardized achievement tests in English, Chinese, mathematics, science, geography, and history in grade seven (T1), grade eight (T2), and grade nine (T3). Achievement tests were administered to all students in the language of instruction in which the student studied the particular subject (i.e., students studying a subject in Chinese completed the test in Chinese, whereas students studying a subject in English completed the test in English), but were otherwise identical. The tests for the six school subjects for each of the first three years of high school were constructed by working parties brought together by the Hong Kong Department of Education with representation from the Advisory Inspectorate Division, the Curriculum Development Institute, and Educational Research Section. In late May or early June of each year, all students who were present on the day that the tests were administered completed achievement tests according to a modified random matrix sampling design in two testing sessions conducted on the same day. In the first testing session, each student was randomly assigned an achievement test in one of three core

subjects: Chinese, English, and mathematics.[3] In the second test session that was conducted on the same day, students were randomly assigned to take a test in one of three additional subjects: geography, history, and science.[4]

*Language of Instruction*

The main independent variable in this study was language of instruction. At the start of the study, the Hong Kong Department of Education broadly classified schools according to their language of instruction (English, Chinese, and mixed Chinese/English) and students' academic ability level. This was done using a combination of public documents and the knowledge of academic subject inspectors and other officers who worked for the Department of Education. For purposes of this study, however, the Hong Kong Department of Education subsequently collected more detailed information about the language of instruction by surveying all students in participating schools at T3, when students were in grade nine. Each student completed a survey about the use of English in the school and in particular school subjects (other than Chinese). For each of the specific school subjects (English, mathematics, science, geography, history, social studies), students responded to four questions: language used in tests and examinations, language used for homework assignments, language used for textbooks, and the actual language used by the teacher. The first three items were measured on a three-point response scale (1 = Chinese only; 2 = mixed; 3 = English only), whereas the final question was answered using a 7-point response scale (1 = all Chinese, not a single sentence in English was spoken; 2 = almost all Chinese, with a few sentences of English explanation; 3 = mainly Chinese, but often supplemented with English; 4 = always switching between Chinese and English explanations and terms; 5 = mainly English, but often supplemented with Chinese; 6 = almost all English, with a few sentences of Chinese explanation; 7 = all English, not a single sentence in Chinese was spoken). There were eight additional items referring to use of English in other (nonclassroom) school activities (e.g., school notice boards, sport events, morning assembly, graduation ceremonies, open day, newsletters), again using a 7-point response scale.

In order to explore the dimensionality of the responses, separate exploratory factor analyses were conducted, one based on responses by individual students and one based on school-average responses (i.e., each of the 56 cases was the mean response to each item by all students within the particular school). Both analyses demonstrated three distinct language-of-instruction factors: Instruction in English (in all classes other than English, keeping in mind that Chinese was not included in the survey), use of English in English Classes, and English Ethos (use of English in nonclassroom aspects of the school). Based on these preliminary factor analyses and for purposes of this study, we constructed three scores (English Instruction, English in English Classes, and English Ethos) to represent language of instruction for each school. These scores varied along a "primarily Chinese" to a "primarily English" continuum. The three

---

[3] For example, the likelihood of a student completing a math exam in any one year was 1/3 and the likelihood of any one student completing a math exam in all three years was 1/27 = 1/3 x 1/3 x 1/3.

[4] This randomization procedure worked effectively for the three core subjects in the first testing session, in that groups of students taking each test did not differ significantly from each other on the pretest achievement score (Ach0) common to all students. The randomization procedure was not fully effective for the second set of tests in that some schools did not offer both history and geography so that only two out of the three (geography, history, science) achievement tests were used in the second test session for these schools. Thus, for example, if history was not offered in a particular school, then each student in that school was randomly assigned to complete either the science test or the geography test, but no students were assigned to complete the history test. This strategy resulted in a somewhat higher proportion of science tests in that science was offered by all schools. The group of students taking the science test, however, had significantly lower Ach0 scores than did the groups of students taking the history or geography tests. However, because analyses were conducted for each subject separately, this potential problem is not a critical issue.

language-of-instruction variables were highly correlated: $r = .66$ ($p < .001$) for English Instruction and English in English Classes; $r = .49$ ($p < .001$) for English Instruction and English Ethos; and $r = .34$ ($p<.001$) for English Instruction in English Classes and English Ethos.

*Statistical Analysis*

Educational research typically involves hierarchically ordered data in which there are multiple units of analysis. In particular, students are typically nested within classrooms or schools. It is usually inappropriate to treat responses by individual students as if they are a random sample without regard to schools because students within the same school are typically more similar to each other than they are to students from different schools (a violation of the independence of statistical tests that do not take the multiple levels – student and school – into account). Also, if questions of interest involve both individual students and schools, then it is more appropriate to conduct multilevel analyses that allow the researcher to simultaneously evaluate results at both units of analysis than to consider only one of the potential units of analysis. Moreover, relations observed at one level of analysis might not bear any straightforward connection to relations observed at another level. Multilevel analyses allow researchers to simultaneously consider multiple units of analysis within the same analysis. A detailed presentation of the conduct of multilevel modeling (also referred to as hierarchical linear modeling) is available elsewhere (e.g., Bryk & Raudenbush, 1992; Goldstein, 1995; Goldstein et al., 1998; Raudenbush & Bryk, 1988). In the present investigation, statistical analyses consisted of multilevel analyses conducted with the commercially available MLwiN (Goldstein et al., 1998) statistical package.

Multilevel growth modeling (see Bryk & Raudenbush, 1992; Goldstein, 1995) is a statistical approach in which growth in student achievement over time can be compared within and across different school types. In the present investigation, the dependent variable is achievement in each of six school subjects over the T1-T3 period. Pretest (T0) achievement measures are used as a covariates to correct for initial student differences (see Plewis, 1996a, 1996b) and to evaluate how growth in achievement varies as a function of pretest (T0) achievement, language of instruction, and their interaction (i.e., aptitude-treatment interactions). More specifically, we fit a three-level growth model in which the three levels were time (the occasion of the achievement test score: T1, T2, T3), student ($n = 12,784$), and school ($n = 56$). This model is described in more detail in the appendix.

Multilevel growth modeling offers an attractive approach to the analysis of longitudinal data, as growth trends are allowed to vary for each student. The growth modeling approach does not require all individuals to have the same number of data points over time and provides an efficient approach to the common problem of missing data in longitudinal research. That is, in the same way that the multilevel modeling procedure can handle varying numbers of students in different schools – assuming that the sample of students is a representative sample of the school – such a procedure can incorporate varying numbers of data points for each person, provided that the points are a representative sample of student's achievement. Goldstein (1995) emphasized in particular the appropriateness of this approach for repeated measures of data and, more generally, for multivariate data in studies in which "measurements are missing by design rather than at random" as "in certain kinds of educational assessments, known as matrix sample designs" (p. 7). In the present study, for example, each student completed only two randomly assigned achievement tests from the six achievement tests that were considered so that scores on the other four tests were "missing" in accordance with the matrix sample design of the study. Hence, this multilevel growth modeling approach is ideally suited to the present investigation.[5]

[5] With multilevel analysis, it is possible to conduct a multivariate analysis that incorporates all school subjects. In the present application this approach was not possible because a particular student on any one occasion completed only one of the three tests administered in the first session (mathematics, English, and Chinese) and only one of

Particularly in multilevel models, data transformations facilitate interpretations. Following Marsh and Rowe (1996; also see Aiken & West, 1991; Bryk & Raudenbush, 1992), we began by standardizing (z-scoring) all variables to have M = 0, SD = 1 across the entire sample.[6] Product terms were the product of individual (z-score) standardized variables (and were not re-standardized). Coefficients for the linear growth components were standardized so that the squared coefficients summed to 1.0.

Due to the nature of some of the variables, and potential problems associated with multicollinearity, we also "residualized" several of the predictor scores. We used five pretest variables: pretest achievement (Ach0), the basis of student selection into high schools, was the primary pretest variable; we also considered standardized achievement tests scores (verbal and mathematics) and school-based performance measures (in English and Chinese). Each of these additional test scores, however, is substantially correlated with Ach0 and was used in the construction of the variable Ach0. That is, Ach0 was an aggregate of school-based performance measures including the English and Chinese school-based performance measures, and the standardized achievement tests were used to moderate scores from school to school. The four additional achievement pretest scores were "residualized" by partialling out the effects of Ach0. Thus, for example, the effect of the residualized pretest English achievement represents the effect of English achievement that is independent of Ach0. This is analogous to a hierarchical approach in which all variance that can be explained by Ach0 is attributed to this variable and only variance that can be explained uniquely by pretest English is attributed to that variable.

The main language of instruction variable was English Instruction (EInst), but we were also interested in the additional effects of use of English in English Classes (EEng) and English Ethos (EEthos). Because these variables were highly correlated, we partialled out the effects of the English Instruction from English in English Classes and English Ethos. Thus, for example, the effect of English in English Classes represents the effect of this variable independent of English Instruction.

---

the three tests administered in the second session (history, geography, and science) so that correlations among the achievement scores within each set could not be estimated (i.e., no students completed both English and mathematics tests at T1).

[6] Goldstein (1995) observes that for educational achievement data on the same individuals over time, it is common to standardize the measures so that they have the same population distribution at each occasion, noting that whereas no trends in the means or variances over time can be estimated, between-individual variation can be estimated and evaluated with multilevel models like those used here. In our study, because the actual items on each achievement test differed from year to year, we could not compare the absolute scores from one year to the next. Thus, for example, we could not say that English or mathematics achievement – averaged across all students – increased over the three years of the study. Instead, as is common in educational research, achievement scores were standardized (mean = 0, SD = 1) separately for each occasion. As emphasized by Goldstein (1995), however, we could determine relative changes over time for any one student or for students in any particular school. Thus, for example, if students in a particular school had an average mathematics achievement z-score of zero (i.e., were average) for the first year of the study (T1), + .25 (i.e. .25 SD above the mean) at T2, and +.50 at T3, then we could claim that a linear increase occurred in achievement over the three-year period (relative to the scores of the entire sample of students in the study who were broadly representative of the Hong Kong population). Not being able to specify absolute growth was not a limitation in this study because absolute growth without reference to some appropriate standard of comparison is typically not very useful. For example, knowing that students in a particular school on average were able to answer correctly one more mathematics item at T2 than T1 would not be useful unless we knew how this finding compared with performances by other students at other schools (or some normative comparison group). Because of this problem, scores on standardized achievement tests are typically normed separately for students at each grade level (year in school). For example, IQ scores are normed (mean = 100, SD = 15) separately for students of different ages so that changes over time for a given student can only be used to infer growth relative to a normative sample. Hence, the focus of the present investigation was on relative levels of student achievement, relative growth in student achievement, and how this relative achievement varied with different characteristics of individual students and the schools that they attended (see related discussions by Goldstein, 1995; Goldstein et al., 1998; Marsh & Grayson, 1994).

BEST COPY AVAILABLE

In the present investigation, as in all longitudinal studies, the appropriate handling of missing data was an important issue. Missing values in dependent variables (i.e., achievement scores in different school subjects administered at T1, T2, and T3) due to the matrix sampling design did not pose a problem because of the multilevel approach to growth modeling used in this analysis, as explained earlier. All students in the study had pretest achievement (Ach0) scores. Missing values for the other pretest scores were imputed by using scores predicted by Ach0. Since these scores were all completely standardized residuals based on prediction from Ach0, this procedure resulted in assigning all missing values a residual value of 0, indicating no difference from the score predicted by the pretest achievement score. Because the language-of-instruction variables were all measured at the school level, no missing values existed for any of these variables.

## Results

### *Preliminary Results*

The correlations between pretest achievement, language of instruction, and post-test achievement are presented in Table 1. Although the total sample size was 12,784, the combination of the matrix sampling and missing data meant that the number of achievement scores for a given subject at a particular occasion was much smaller. The residualization strategy is also evident from these correlations in that pretest achievement (Ach0 in Table 1) is uncorrelated with residual scores for each of the remaining (residualized) pretest scores, even though the original (unresidualized) scores for all these variables were highly correlated. Similarly, general English Instruction (EInst in Table 1) was uncorrelated with the remaining two (residualized) English-language scores.

### *Pretest-Post-test Correlations*

The correlations between pretest and post-test achievements support the construct validity of interpretations of the test scores (Table 1). The unresidualized total pretest achievement score was substantially and significantly correlated with all post-test achievement scores and the sizes of the correlations were nearly as high at T3 (three years later) as at T1 (due to the very large sample size in this study, almost all correlations are "statistically" significant and so we will focus on whether the size is substantial rather than statistical significance per se). The patterns of correlations between the specific residualized pre-test achievement and post-test achievement scores, however, varied substantially (i.e., varied to an extent that was of practial significance as opposed to statistical significance) depending on the particular school subject. For example, the pretest Chinese grades and test scores (Chin0 and Verb0) were moderately correlated with subsequent achievement in Chinese (e.g., correlation coefficients of .21and .28 for T1 Chinese achievement, see Table 1), recalling that these correlations represent contributions beyond what was explained by the total pretest achievement score. Similarly, the residualized pretest numeric scores were moderately correlated with subsequent achievement in mathematics (e.g., r of .27 for T1 mathematics achievement, see Table 1) and, to a lesser extent, science (e.g., r of .17 for T1 mathematics achievement, see Table 1). Also, the residualized pretest English scores were moderately correlated with subsequent performance in English (e.g., r of .25 for T1 English achievement, see Table 1). Not surprisingly, there is an overall pattern of slightly stronger relations between pretest (T0) achievement and T1 achievement than between T0 achievement and T3 achievement, but the differences are small.

### *Language-of-Instruction Correlations*

All three of the language-of-instruction scores (EInst, EEng, Ethos) were empirically derived from student responses to the survey on the use of English collected at T3, whereas one (English language of

instruction, ELLOI, in Table 1) is based on an a priori classification provided by the Department of Education at T1. The variable ELLOI correlated .91 with the empirically derived English Instruction scores (EInst) and the pattern of correlations with other variables was very similar for these two general language-of-instruction variables. In particular, the correlation between ELLOI and each of the other variables in Table 1 is nearly the same as the corresponding correlation with EInst [e.g., r(Ach0, ELLOI) = .45, r(Ach0, EInst) = .47]. Hence, the a priori classification of schools made by the Department of Education at T1 based on public records and local knowledge (ELLOI) agreed remarkably well with the empirically derived English Instruction score collected at T3 (EInst). This provides extremely strong support for the construct validity of the language-of-instruction variables. Since the other two empirical language-of-instruction variables were residual scores, they were nearly uncorrelated with both the general scores.

Pre-test achievement was substantially related to language of instruction. For example, pre-test achievement correlates .47 with general English Instruction and .47 with the (residualized) English in English Classes. However, because the pre-test achievement scores were based on achievement before students entered high schools, we may attribute some of this correlation to school selection processes: students with higher pre-test scores tend to attend the more prestigious English-language high schools in Hong Kong. Interestingly, language-of-instruction variables (ELLOI, Einst, EEng) were almost uncorrelated with the residualized pretest English scores (Eng0 Resid, see Table 1). Thus, although students in English-language high schools were much brighter than average, they were not particularly proficient in English, beyond what would be expected in terms of their general achievement. Hence, interpretations must account for these large initial pretest differences and relations that do not control for initial differences (i.e., the correlations as presented in Table 1) should be interpreted cautiously.

Correlations between the English Instruction and post-test achievement scores varied widely for different school subjects. For all three time points (see Table 1), the correlations were substantially positive for English, Chinese, and mathematics, but are substantially negative for history, geography, and science. However, these correlations do not control for the higher pre-test achievement of students in English-language high schools. Thus, much of the apparent advantage of attending English-language high schools for English, Chinese, and mathematics was due to preexisting differences and not to language of instruction. For example, the positive correlations between English Instruction (EInst in Table 1) and achievement in English, Chinese, and mathematics were consistently smaller than the strong positive correlation between English Instruction and pretest achievement (Ach0) science. Even more dramatic, the apparent disadvantage of attending English-language high schools for geography, history, and science was based on uncorrected scores that did not take into account the fact that these students were much brighter than average before entering these high schools. Despite the fact that these students were much brighter than average, their scores in these three subjects were much lower than average. Controlling for preexisting differences significantly increased these negative effects. Additionally, the negative correlations associated with attending English-language schools declined somewhat over time for science and history – suggesting that the negative effects associated with attending English language schools might become smaller over time. In summary, these results show that students attending English-language high schools score below average in geography, history, and science, even though these students were initially more able and should have performed well above average based on their pretest achievement levels.

In marked contrast to general English Instruction, the use of English in English Classes (EEng) was consistently positively correlated with achievement in all school subjects (Table 1). Because use of English in English Classes was a residual score (controlling for general English Instruction), these correlations indicate that students in English classes where the emphasis on English was stronger than

BEST COPY AVAILABLE

expected on the basis of the level of English Instruction did better in all school subjects than students in schools where the emphasis on English in English classes was not strong. Two features, however, complicate interpretations of this relation. First, much of this difference was due to preexisting (pretest achievement) differences. Second, a strong emphasis on English in English Classes would likely be advantageous in learning history, for example, if students were taught history in English, but might not be advantageous if students were taught history in Chinese. The residualized English Ethos of the school (use of English in nonclassroom settings, EEthos) was uncorrelated with each post-test achievement score, suggesting that English Ethos was not an important variable in predicting achievement scores.

In summary, these preliminary analyses provide possible answers to our overarching question and at least some of our more specific research questions. In particular, the effects of instruction in English have negative effects for at least some school subjects (particularly nonlanguage subjects), but these effects seem to vary substantially depending on the particular school subject. Also, the effects of instruction in English appear to be reasonably stable over time. These results are important because they provide preliminary results that are not complicated by adjusting for prior differences in achievement and do not involve complicated statistical analyses. They are also limited, however, for these same reasons. Thus, we now turn to the statistically stronger and more appropriate longitudinal multilevel models of these same data.

*Longitudinal Multilevel Analyses*

Separate analyses were conducted for each school subject (Table 2; also see Figures 1 - 6). For each subject, five models were fit. Model 1 (a variance components model) included no predictor variables, but provided a baseline for how much variance in scores could be attributed to differences in school, differences between students, and differences within students over time. In a series of Models 2-5, new variables were added one step at a time according to an a priori sequence to explicate these results. In Model 2, pretest achievement variables were included to determine how much variation between schools could be explained by preexisting differences. In Model 3, language-of-instruction variables were added to evaluate their effects on subsequent achievement. In Model 4, interactions involving language of instruction were included to see if the effects of language of instruction varied with pretest levels of achievement. Finally, in Model 5, the growth and stability of the effects over time were evaluated. In addition to the fixed effects associated with these predictor variables, we assessed the extent to which variation from school to school could be explained by controlling for the variables included in each of these models.

*Mathematics*

Post-test mathematics achievement (see Model 5 for mathematics in Table 2) was substantially related to pre-test achievement (Ach0) and to the mathematics component of the pre-test achievement test (Mth0). Our analysis indicated that English Instruction had a small negative effect on post-test mathematics achievement, but that English in English Classes had a small positive effect (see Model 5 in Table 2 and Figure 1 ). The small negative effect of time overall was not substantively important because different tests were used on each occasion.[7] Overall, the small negative effect of English Instruction in general did

[7] The number of students completing achievement tests declined slightly in each year of the study (11,528 in T1, 11,045 at T2, and 10,900 at T3), representing a combination of absence on the particular day the achievement tests were administered and students who withdrew or changed schools. Furthermore, the pretest achievement scores (Ach0) that were available for all students were slightly higher for students who completed tests at T2 than at T1 (+.047 SD at T2 vs. +.014 SD at T1) and higher still at T3 (+.073 SD). Hence, students who were in the study at T3 tended to be slightly brighter (based on the pretest achievement score) than those who were not. This explains



BEST COPY AVAILABLE

not vary with pretest achievement, but the small three-way interaction involving time suggests that over time this negative effect became somewhat smaller for initially brighter students. Inspection of the variance components for the five models of post-test mathematics achievement suggests that the substantial school-to-school variation in mathematics achievement was largely explained by the pretest variables – the variance component (L3VSch under random effects in Table 1) of .39 in Model 1 dropped to .04 with the addition of pretest variables in Model 2, and decreased only slightly with the addition of language-of-instruction variables in subsequent models. Thus, it is not surprising that school-type differences in relation to language of instruction were not large.

*Chinese*

Post-test Chinese achievement was substantially related to pretest achievement (Ach0) and pretest Chinese achievement (verb0 and Chin0). English Instruction had a small positive effect, but interestingly, English in English Classes had a slightly more positive effect on post-test Chinese achievement (see Figure 2 ). As was the case with mathematics, the small positive effect of English Instruction did not vary with pretest achievement, but the small three-way interaction involving time suggests that over time this positive effect became somewhat smaller for initially brighter students. Inspection of the variance components for post-test Chinese achievement suggests that the substantial school-to-school variation (.46 in Model 1) was largely explained by the pretest variables (the variance component was reduced to .04 in Model 2), although the inclusion of language of instruction resulted in a further drop in residual variance due to school-to-school variation (the variance component was .01 in Models 3-5).

*English*

Post-test English achievement was substantially related to pretest achievement (Ach0), prior English achievement and, to a lesser extent, the verbal (Chinese) pretest achievement. As shown in Figure 3 , English Instruction had a substantial positive effect, but not surprisingly, English in English Classes had an even more positive effect on post-test English achievement . Overall, the positive effect of English Instruction was somewhat greater for initially brighter students (the Ach0 x EInst effect in Table 2) and increased slightly with time (the Time x EInst effect in Table 2). Although initially brighter students did substantially better, this advantage declined slightly over time (Time x Ach0 effect in Table 2). The small three-way interaction (Time x EInst x EEng in Table 2) suggests that being in a school that had a strong general emphasis on English Instruction and a strong emphasis on English in English classes had an initial positive effect that declined over time. Inspection of the variance components for post-test English achievement suggests that the substantial school-to-school variation (.60 in Model 1) was explained in large part by the pretest variables (the variance component was reduced to .11 in Model 2). The inclusion of language-of-instruction variables, however, resulted in a further drop in the school variance component (the variance component was .03 in Models 3-5).

*History*

---

why there is a slight tendency for achievement tests to decline in the multilevel growth models (the linear effect of time is slightly negative), even though the scores were standardized separately at each occasion. Thus, for example, a student at the mean achievement test score at T1 would tend to be slightly below the mean at T3 (in relation to the slightly brighter cohort of students at T3 compared to those at T1). In contrast, students who completed test scores on all three occasions had scores approximately +.07 SD on the pretest achievement score and total test score (averaged across the three occasions). Because this small effect of time in the multilevel models is not substantively important, it is not discussed further.

Post-test history achievement was substantially related to pre-test achievement (Ach0) and, to a much lesser extent, the pretest Chinese (verb0 and Chin0) and English achievement. Because the effects of language of instruction for history were similar to those for geography and science, we describe the history results in somewhat greater detail. Model 5 (Table 2) demonstrates that English Instruction had a very large negative effect on history achievement. Thus, for example, the negative effect of English Instruction (ß=-.67) was more negative than the positive effect of pre-test achievement (ß=.44). Students taught history in English were strongly disadvantaged relative to students who were taught in Chinese. However, our analyses demonstrate that English in English Classes had a substantial positive effect on history achievement and that this positive effect was particularly large in schools with more English Instruction in general. These main and interaction effects are illustrated in Figure 4 . For students in schools where the language of instruction was primarily in Chinese (-1.5 SD on English Instruction), the mean post-test history score was about one standard deviation above the mean of history achievement, and English in English Classes had no effect. For students who were taught history primarily in English-language schools (+1.5 SD on English Instruction), the mean history achievement was about one SD below the mean of history achievement. Here, however, the emphasis on English in English Classes made a big difference. Students who were in schools with high scores (+1.5 SD) in both English Instruction and English in English Classes had post-test history achievement scores that were about average, but students in schools with a high score in English Instruction but a low score in English in English Classes did much poorer. Thus, whereas students were strongly disadvantaged by being taught history in English, this effect could be slightly offset by attending schools with a particularly strong emphasis on English in English classes.

The very negative effects of English Instruction on student achievement were balanced somewhat if the student possessed strong prior English skills (the Eng0 x EInst interaction). As shown in Figure 5 , the negative effect of being taught history in English was evident for all levels of initial English achievement, but the effects were somewhat smaller for students with initially strong English skills (high [+1.5 SD] English Pretest Skills in Figure 5 ) compared with students with initially weak English skills (low [-1.5 SD] English Pretest Skills in Figure 5 ). Over time, the very negative effects of English Instruction declined somewhat (the time x EInst interaction). As shown in Figure 6 , the very negative effect of English Instruction was slightly larger at T1 (grade seven, the first year of high school) than at T3 (grade nine). Inspection of the variance components for post-test history achievement suggests that – unlike the models of post-test English, Chinese, and mathematics – the substantial school-to-school variation post-test history achievement (.61 in Model 1) was not substantially eliminated by controlling for pretest variables (the variance component remained .50 in Model 2). Only with the inclusion of language-of-instruction variables was there a substantial drop in the school variance component (the variance component was reduced to .07 in Model 3). However, the addition of the aptitude-treatment interactions in Model 4 and interactions with time in Model 5 also resulted in further reductions (see Table 2) in school-to-school variation. In marked contrast to mathematics, English, and Chinese, these variance components indicate that language-of-instruction school types did make a substantial difference in school-to-school variation in all subsequent history achievement, although these effects were moderated to some extent by pretest aptitudes and time.

*Geography*

Post-test geography achievement was strongly related to pretest achievement (Ach0) and, to a much lesser extent, the pretest Chinese (verb0 and Chin0) and mathematics (Mth0) achievement. A general emphasis on English had a very large negative effect on geography achievement, but a strong English emphasis in English classes had a large positive effect – particularly when there was also a strong general emphasis

on English. (This was a similar pattern of results to that described for history achievement in Figure 4 and, thus, is not described again in detail). The very negative effect of the general emphasis on English was offset somewhat by having strong prior English skills (the Eng0 x EInst interaction; see also related results for history in Figure 5 ). The negative effects of the general emphasis on English did not, however, vary with time. Inspection of the variance components for post-test geography achievement suggests that the substantial school-to-school variation (.50 in Model 1) was not substantially eliminated by controlling for pretest variables (the variance component was .38 in Model 2). Only when language-of-instruction variables were included in Model 3 was there a substantial drop in the school variance component (.04 in Model 3), although the addition of the aptitude-treatment interactions in Model 4 also resulted in a small further reduction in school level residual variance. As with history, these variance components indicate that language-of-instruction school types did make a substantial difference in all subsequent geography achievement.

*Science*

Post-test science achievement was strongly related to pretest achievement (Ach0) and, to much lesser extents, pretest Chinese (verb0 and Chin0) and mathematics (Mth0) achievement. A general emphasis on English instruction had a very large negative effect on science achievement, but a strong English emphasis in English classes had a large positive effect. The effect of a strong emphasis on English in English classes was particularly positive when there was also a strong curricular emphasis on English (EInst x EEng). Furthermore, the corresponding three-way interaction (Time x EInst x EEng) indicates that this positive effect increased somewhat over time. The very negative effects of the general emphasis on English were offset somewhat by having strong prior English skills (the Eng0 x EInst interaction). The negative effects of the general emphasis on English also declined somewhat with time (Time x EInst interaction). The substantial school-level variance components for science achievement suggest that school-to-school variation (.34 in Model 1) was not reduced by controlling for pretest variables (the variance component was .39 in Model 2). Only when the language-of-instruction variables were included in Model 3 did a substantial drop in the school variance component occur (the estimated variance dropped to .07 in Model 3). However, the addition of the aptitude-treatment interactions in Models 4 and 5 also resulted in small further reductions in school level residual variance. As observed with history achievement, these variance components indicate that language-of-instruction school types did make a substantial difference in all subsequent science achievement.

Discussion

*Do Effects of Instruction in English Vary for Different School Subjects?*

The results provide a dramatic affirmative response to this first research question. For two subjects, Chinese and, particularly, English, the effects of English Instruction were moderately positive; for one, mathematics, there were small negative effects; and for three subjects, history, geography, and science, the effects were extremely negative. The positive effects of Instruction in English on post-test English achievement were not surprising, which is why public advocates (and parents) argue in favor of English-language high schools. Although the effects for Chinese were small, it is interesting that the effects were positive and not negative. This finding suggests that learning a L2 can benefit L1 achievement and is consistent with earlier results for immersion studies and predictions based on the interdependence hypothesis from Cummins's 1979 model (see also Francis, . 1999).

The most important findings, however, were the very strong negative effects of Instruction in English on history, geography, and science. For each of these three subjects, the negative effects of Instruction in English were about as large, or larger, than the positive effects of pretest achievement. The apparent

BEST COPY AVAILABLE

similarity in these three subjects is that each involves a relatively new content area for students first entering grade seven and requires students to learn new terminology in order to understand the conceptual underpinnings of these subjects. When students are forced to do this in an L2 (English) that is not already well-mastered, students must place undue attention on mastering basic terminology that may preclude gaining a deeper conceptual understanding of these subjects, active participation in classroom discussion, and even reading the textbook that is also in English. Following this reasoning, because Chinese classes are taught in Chinese and English courses are taught in English, it is not too surprising that the language of instruction for other school subjects in the school did not have any negative effects on Chinese and English achievement. For mathematics, the English Instruction effects were negative, but much smaller than for history, geography, and science. Teaching in mathematics, however, is based largely on a symbolic terminology that may not be so dependent on the language of instruction and may have already been more adequately mastered prior to grade seven relative to the other (nonlanguage) subjects considered here. We suggest that this has to do with the development of mathematics knowledge in general and is not specific to instruction in Hong Kong. This suggestion is supported by the similar pattern of results presented by Willig (1985), based on her review of U.S. research.

*Do Effects of Instruction in English Vary With Pretest Academic and English Competency?*

From both theoretical and policy perspectives, it is important to determine whether the negative effects of Instruction in English on history, geography, and science vary depending on the initial aptitudes of the students (i.e., whether or not there are aptitude-treatment interactions). Fortunately, the quality of our pretest measures are exceptionally good in that they are comprehensive, reliable, and highly correlated with achievement in subsequent years. Contrary to expectations (based on predictions from Cummins 1979 model and from the rationale for current policy in Hong Kong about the allocation of students to English-language schools), the disadvantages associated with being taught in English were not smaller for initially brighter students than for initially less able students (i.e., the EInst x Ach0 interactions were nonsignificant for history, geography, and science, as shown in Table 2). For all three content subjects, however, students who had initially better English skills were somewhat less disadvantaged by instruction in English. The juxtaposition of these two sets of results is important because earlier results (see Table 1) indicated that the allocation into English-language high schools was based primarily on the total pretest achievement scores and not specifically on prior English skills. The results suggest, perhaps, that more emphasis should be placed on prior English skills when assigning students to English-language high schools in order to minimize the negative effects of Instruction in English. Such an educational policy, however, may contradict the immersion philosophy of starting with students who have limited (or no) L2 proficiency (Swain & Johnson, 1997) and may have undesirable side effects of placing even more emphasis on English, possibly devaluing Chinese, and giving the appearance of a more elitist program in that only students with stronger initial English can study in the prestigious English-language high schools (Gibbons, 1989).

*What Are the Effects of English in English Classes and English Ethos?*

To the extent that English Ethos accurately reflected the use of English in nonclassroom activities, our results indicated that English Ethos did not contribute significantly to achievement in any of the six school subjects. In contrast, the effects of English in English Classes had positive effects on achievement in all six subjects, even after controlling for pretest achievement and Instruction in English in general. It is not surprising, of course, that a stronger emphasis on English in English Classes leads to better achievement in English and, perhaps, Chinese (as predicted by the Cummins 1979 model and reflecting the rationale of immersion programs).

More interesting is the question of why a stronger emphasis on English in English Classes has positive effects on achievement in history, geography, science, and, to a lesser extent, mathematics. Apparently, these positive effects of strong English classes extend to other classes taught in English, such that those students are less disadvantaged by Instruction in English. For history, geography, and science, there were interactions between the English in English Classes and the Instruction in English in general. For example, as illustrated in Figure 4 for history, the emphasis on English in English classes had no effect on achievement when the history class was taught primarily in Chinese (i.e., General English Use is low [-1.5 SD] on the left side of the graph) but had a substantial positive effect when history was taught primarily in English (i.e., General English Use is high [+1.5 SD] on the right side of the graph). Even in mathematics, where the effects of language of instruction were much smaller, the direction of this nonsignificant interaction effect (.036, SE = .020) was positive. From a policy perspective, these results are very important, suggesting that having a particularly strong emphasis on English in English Classes can offset some of the negative effects of Instruction in English in nonlanguage subjects.

*Do Effects of Language of Instruction Vary Over Time?*

Contrary to expectations, the effects of language of instruction did not vary substantially over time. In particular, we anticipated that the Instruction in English effects on history, geography, science, and mathematics might be relatively more negative in the first year of high school when students were first introduced to Instruction in English, but might become smaller as students became more accustomed to Instruction in English and acquired better English-language skills. There was weak support for these expectations for history and, to a lesser extent, science, but not for geography or mathematics. Whereas this support was strongest for history, the results (see Figure 6 ) demonstrate that the reduction in the negative effects over time was not substantial. Whereas the achievement differences between schools taught primarily in English and primarily in Chinese were somewhat smaller at T3 than T1, the differences were not large. Although other interactions involving time existed, these effects were small and not consistent across different school subjects.

In summary, Hong Kong high school students were very disadvantaged by Instruction in English in geography, history, science, and, to a lesser extent, mathematics. The size of this disadvantage was reasonably consistent across the first three years of high school. Although the size of this disadvantage did not vary much with initial achievement levels in general, the disadvantage of Instruction in English was somewhat smaller for students who initially had better English-language skills. Furthermore, these large negative effects of Instruction in English in English-language schools were offset to a limited extent by a strong emphasis on English in English Classes.

Implications for Theory and Generalizations Based on Previous Research

The Canadian immersion studies (Lambert, 1992) showed that instruction in L2 (French) for English Canadians with little or no prior L2 proficiency had positive effects on subsequent L2 proficiency, but also on achievement in L1 (English) and some other school subjects. In marked contrast, Willig (1985) concluded that for students with limited L2 (English) proficiency there were consistently positive effects of teaching students in their L1 (in bilingual transition programs) rather than L2, but she excluded the Canadian immersion studies. Cummins (1979) provided a theoretical model of L2 acquisition that seemed to be consistent with these seemingly contradictory results. In particular, he argued that students would only realize the benefits of bilingualism if they were sufficiently competent (i.e., above a proficiency threshold) in both languages, when the ongoing development of L1 was reinforced outside of school (e.g., it was the dominant language in the particular society), and, perhaps, if students were motivated to learn and appreciate both languages. Whereas this theoretical position seemed consistent with both the original

BEST COPY AVAILABLE

Canadian immersion studies and the bilingual transition studies, subsequent immersion studies suggested that the benefits of immersion generalized more broadly than might be expected from Cummins's theoretical model. Much of this research, however, was based on early immersion programs.

Reviews of both immersion and bilingual transitional paradigms have focused almost exclusively on language proficiency, and little attention has been given to achievement in nonlanguage classes. Thus, for example, Willig (1985) reported very few studies that considered social studies achievement and reported no studies that evaluated science achievement. Given the strong applied-policy orientation of most research in this area, this is a shocking omission. Epitomizing this perspective in a discussion of the Hong Kong context from the perspective of immersion research, Johnson (1997) suggested that if the goal of immersion is to create Hong Kong students who are highly proficient in English, then achievement in nonlanguage subjects might be irrelevant. In marked contrast to this apparent disregard for achievement in nonlanguage subjects, the Hong Kong Education Commission (1990) specified that English should only be used if a student's *overall* cognitive and academic development benefited.

We reject Johnson's perspective, which seems to permeate immersion research, and interpret the results of our study – with some qualifications – as largely contradicting the implications of previous immersion studies and, apparently, predictions from Cummins's interaction theory of second language acquisition. Overall, the effects of the immersion program were negative – not positive. These conclusions are important because our study seems to match closely the main characteristics of the prototypical immersion study (e.g., Swain & Johnson, 1997) and seem to satisfy all of the conditions that Cummins indicates are important to achieve the positive benefits of bilingualism. Yet, overall, the results of this study suggest that immersing high school students into L2 instruction has very negative effects. An important qualification, of course, is that the late-immersion program considered here did have small positive effects on language achievement, and this has been the primary criterion used in most other research. From this overly narrow perspective, it may be possible to argue that the effects are consistent with previous immersion studies. When compared to the overwhelmingly negative effects of immersion for the nonlanguage subjects, however, we consider that our large-scale quasi-experimental study – one of the largest late-immersion studies ever conducted – was a failure in terms of providing academic benefits for Hong Kong students, as well as supporting predictions based on previous immersion research and Cummins's theory.

Why does our research contradict generalizations based on previous immersion studies and theory? The two most likely suggestions seem to be our emphases on (a) achievement in nonlanguage as well as language subjects, and (b) a late- rather than an early-immersion program. In most previous research, there has been a remarkable disregard for achievement in nonlanguage subjects, and research has focused on early-immersion programs. Support exists for both these suggestions based on qualifications that Cummins (1979) offers for his model and implications from Willig's (1985) review. Cummins (1979) emphasized that the threshold of L2 competency needed to achieve benefits from immersion might be much higher at higher levels of schooling and that children may need to experience the immersion early in their schooling, when the language demands are sufficiently low that children can gain L2 fluency:

> Thus, in the early grades the lower threshold may involve only a relatively low level of listening comprehension and expressive skills, but – as the curriculum content becomes more symbolic and requires more abstract formal operational thought processes – the child's 'surface' L2 must be translated into deeper levels of 'cognitive competence' in the language (p. 231).

Willig (1985; see also Hakuta & McLaughlin, 1996) offered the related caveat that comprehension of abstract concepts in nonlanguage subjects, such as social studies and science, requires a high level of

language fluency in the language of instruction even though the focus of the subjects is not languages per se. This observation, coupled with the dearth of research on nonlanguage achievement, led her to call for more research using achievement in nonlanguage subjects to evaluate better the effects of bilingual and second-language programs. Hence, in the context of Cummins's interaction model, it may not be possible for students to gain benefits from a late-immersion program unless they have already achieved a high threshold of functional L2 competency prior to the immersion. On this basis, one might argue that our results were consistent with Cummins's (1979) theory in that many students taught in English may not have reached a critical threshold. This argument, however, becomes circular when L1 and L2 proficiency are used both to evaluate whether students attained the desired threshold and to evaluate the predictions that language achievement will improve. Instead, because the conditions in this study fit so well with those that Cummins says should lead to benefits associated with bilingualism, we interpret our results as an important contradiction to his theory. This conclusion, if supported by subsequent research, requires a substantial rethinking of the generalizability of the benefits of immersion programs and, perhaps, bilingualism and second-language acquisition for high school students. The implications of these interpretations argue against the use of a late-immersion strategy in which students with limited L2 proficiency are taught entirely in L2. Some qualifications exist, however, to this overarching conclusion, due to potential limitations in the present investigation and the need for further research.

Potential Limitations and Directions for Further Research

1. In our quasi-experimental design, large pretest achievement differences between students existed that were related to language of instruction. Importantly, however, we had a particularly strong set of pretest covariates to control for these initial differences, and we used particularly powerful statistical tools (the multilevel analyses) to achieve this purpose. Also, even with no correction for the large initial differences, students taught in Chinese (even though their pretest achievement scores were more than one SD below students taught in English) scored significantly higher than students taught in English for history, geography, and science (see Table 1). Hence, at least the direction of these differences seems robust against alternative explanations due to this potential design limitation.

2. An implicit assumption is made that the quality of teaching was equivalent in high schools differing in language of instruction. In particular, an immersion program requires teachers to be highly fluent in the language that they are teaching (e.g., Swain & Johnson, 1997), but in Hong Kong there is a shortage of high school teachers in nonlanguage subjects who are fluent in English. Recognizing this problem, the Education Commission (1995) recommended that schools hire more native English-speaking teachers and introduce minimum language-proficiency standards for teachers. A compromise may occur between employing teachers who are highly fluent in English and teachers who have high levels of subject mastery, such that the quality of instruction might be confounded with the language of instruction. If, for example, teachers cannot teach effectively in L2, classes may be less interesting, with more emphasis on rote learning of factual material and less quality discussion and debate. Because we had no measures of the quality of teaching effectiveness, we cannot pursue this conjecture in the present investigation (see related discussion by Johnson, 1997). The particular pattern of results in our study, however, suggests that this potential problem was not the primary reason for the negative effects of Instruction in English. If quality of instruction was the critical variable, it seems unlikely that (a) the negative effects in mathematics should be so much smaller than those in history, geography, and science, and (b) that the emphasis on English in English classes would have been able to offset the negative results.

3. The strategy of partialling the effect of general Instruction in English from the English in English Classes scores was defensible in terms of determining how additional variance could be explained by the second variable. In effect, all variance in post-test achievement that could be explained either by

BEST COPY AVAILABLE

Instruction in English or by English in English Classes was attributed to Instruction in English. Because Instruction in English and English in English Classes were substantially correlated ($r = .66$), not using this strategy would have resulted in potential problems of multicollinearity. This strategy, however, tended to maximize variance attributable to Instruction in English. For example, when we redid the analyses of English achievement using Instruction in English and English in English Classes without using this partialling strategy, the Instruction in English effect was not statistically significant, whereas the effect of English in English Classes was, of course, approximately the same. Hence, using this strategy, much of the benefit in English achievement associated with attending English-language schools could be explained by the emphasis on English in English Classes. Similarly, for each of the six school subjects, the effects of Instruction in English became more negative (or less positive in the case of English and Chinese achievement) when the unpartialled English-in-English Classes score was used. In this respect, our results provide the most positive perspective on instruction in English and, thus, are conservative in relation to our conclusion that teaching nonlanguage subjects in English disadvantages students. Our results may, however, underestimate the relative advantages of English in English Classes compared to Instruction in English.

4. Because the negative effects of immersion into instruction in English for history, geography, and science were large, it is unlikely that any substantial subgroup of students taught in English were advantaged in these subjects. The Hong Kong Educational Commission (1990) anticipated that that the most able students (suggested to be the top 30% in terms of prior achievement) would be advantaged (or, at least, less disadvantaged) by immersion into instruction in English. However, no support existed for this expectation in that interactions with prior achievement and instruction in English were consistently small and mostly nonsignificant (see Ach0 x EInst interactions in Table 2). Nevertheless, a small number of students with extremely good prior mastery of English (e.g., students who were born overseas and migrated back to Hong Kong or had a native English-speaking parent) might not be disadvantaged when taught in English. Furthermore, because our study indicated that the negative effects of immersion declined somewhat over time, and because previous studies (e.g., Cummins, 1979) suggest that the benefits of bilingualism may take more than three years to materialize, the negative effects may lessen as English proficiency improves during the remaining three years of high school (i.e., grades 10-12, not studied in this analysis). Alternatively, students may require a sufficiently long transition period, spent entirely on learning English to an appropriate threshold of proficiency, prior to starting an English-language high school. Consequently, because of the potential social and economic advantages of being fluent in English in Hong Kong, there may be justification for instruction in English in high schools for students with sufficient English-language skills not to be severely disadvantaged by Instruction in English. To explore this possibility, further research is needed that considers a longer period of time (e.g., all six years of high school rather than only the first three) and, perhaps, focuses more specifically on the assessment of English fluency using oral and vernacular measures as well as paper-and-pencil tests. Alternatively, future research in Hong Kong may need to focus on early-immersion programs like those that seem to have been successful in Canada (e.g., Lambert, 1992). However, exploration of the feasibility of such alternatives will require careful consideration because (a) the community language environment may not be conducive to attaining fluency in English because English is rarely used outside the classroom, (b) there are inadequate numbers of teachers with good English proficiency, particularly in primary and kindergarten levels, (c) it would be difficult to determine who would be most benefited by these early immersion programs or whether such programs should be open to all, and (d) such programs would have the potential of further devaluing Chinese as a language of instruction and creating a preoccupation with English-language skills to the detriment of other school subjects.

Policy Implications

*There should be a consistently strong emphasis on English in English classes.* Not surprisingly, the emphasis on English in English Classes was positive for English achievement. This finding is consistent with the guidelines for schools as advocated by the Hong Kong Education Department. A stronger emphasis on English in English Classes will lead to improvement in English proficiency, which subsequently will have a positive effect on students' learning of other academic subjects, particularly when those subjects are taught in English.

*Student's prior English (L2) language skills should be given greater emphasis in allocating students to English-language high schools.* Our results show the limited power of using general achievement to predict which students should go to high schools with a strong general emphasis on English. Although brighter students do much better in all school subjects, the negative effect of Instruction in English did not vary with initial ability. Prior English skills (Eng0) appear to be more useful, in that the negative effect of learning through English was somewhat smaller for students who are initially more proficient in English.

*With the possible exception of students who are already proficient in English, little justification exists for the current practice of teaching all school subjects, such as history, geography, or science, in English during the first years of high school.* Our results suggest that students need to be much more proficient in English before they start high school. This proficiency might be facilitated by: (a) starting the immersion much earlier (in primary school or at the very beginning of schooling, as in the Canadian immersion schools); (b) providing a sufficiently long transition period, between the end of primary and the start of high school, that is devoted entirely to learning English to an appropriate threshold of proficiency prior to starting an English-language high school; or (c) giving students stronger support in these content subjects (e.g., extra lessons or bilingual tutors who are able to explain the lesson content in Chinese). Furthermore, although the size of the negative effects of Instruction in English may decline somewhat as students progress through high school, the size of the decline is relatively small, at least for the first three years of high school considered here. Consistent with this recommendation, the Hong Kong Education Department issued strong guidelines to use Chinese as the medium of instruction for subjects other than English in schools where most students do not have the necessary English proficiency to benefit from Instruction in English and emphasized that unless sufficiently strong support and remedial help are offered, the problems of learning various academic subjects through English will not automatically go away, even after several years.

References

Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. Newbury Park, CA: Sage.

Baker, K. A., & de Kanter, A. A. (1981). *Effectiveness of bilingual education: A review of the literature*. Washington, DC: U.S. Department of Education, Office of Planning, Budget, and Evaluation.

Bourhis, R. Y. (1990). Social and individual factors in the development of language acquisition: Some models of bilingual proficiency. In B. Harley, P. Allen, J. Cummins, & M. Swain (Eds.), *The development of second language proficiency* (pp. 134-145). Cambridge, England: Cambridge University Press.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models: Applications and data analysis methods*. Newbury Park, CA: Sage.

Collier, V. (1992). A synthesis of studies examining long-term language minority student data on academic achievement. *Bilingual Research Journal, 16*, 187-212.

Cummins, J. (1979). Linguistic interdependence and the educational development of bilingual children. *Review of Educational Research, 49*, 222-251.

Cummins, J. (1986). Empowering minority students: A framework of intervention. *Harvard Educational Review, 56*, 18-36.

Cummins, J. (1991). Conversational and academic language proficiency in bilingual contexts. *Association Internationale de Linguistique Appliquée Review, 8*, 75-89.

Cummins, J. (1996). Bilingual education: What does the research say? In J. Cummins (Ed.), *Negotiating identities: Education for empowerment in a diverse society* (pp. 97-133). Ontario, CA: California Association for Bilingual Education.

Cummins, J., & Swain, M. (1986). *Bilingualism in education: Aspects of theory, research and practice*. London, England: Longman.

Cummins, J., Harley, B., Swain, M., & Allen, P. (1990). Social and individual factors in the development of bilingual proficiency. In B. Harley, P. Allen, J. Cummins, & M. Swain (Eds.), *The development of second language proficiency* (pp. 119-1133). Cambridge, England: Cambridge University Press.

d'Anglejan, A. (1990). The role of context and age in the development of bilingual proficiency. In B. Harley, P. Allen, J. Cummins, & M. Swain (Eds.), *The development of second language proficiency* (pp. 146-157). Cambridge, England: Cambridge University Press.

Duff, P. A. (1997). Immersion in Hungary: An ELF experiment. In R. K. Johnson & M. Swain (Eds.), *Immersion education: International perspectives* (pp 19-43). Cambridge, England: Cambridge University Press.

Education Commission, (1990 November). *The curriculum and behavioural problems in schools* (Education Commission Report No. 4) Hong Kong: Government Printer.

Education Commission, (1995 December) *Enhancing language proficiency: A comprehensive strategy* (Education Commission Report No. 6) Hong Kong: Government Printer.

Francis, W. S. (1999). Cognitive integration of language and memory in bilinguals: Semantic representation. *Psychological Bulletin, 125*, 193-222.

Garcia, E. E. (1993). Language, culture and education. In L. Darling-Hammond (Ed.), *Review of research in education 1993, vol. 19* (pp. 51-98). Washington, DC: American Educational Research Association.

Gardner, R. C. (1985). *Social psychology and second language learning*. London: Edward Arnold.

Genesee, F. (1978). Is there an optimal age for starting second language instruction? *McGill Journal of Education, 13*, 145-154.

BEST COPY AVAILABLE

Genesee, F. (1984). Beyond bilingualism: Social psychological studies of French immersion programs in Canada. *Canadian Journal of Behavioural Science, 16*, 338-352.

Genesee, F. (1985). Second language learning through immersion: A review of U.S. programs. *Review of Educational Research, 55*, 541-561.

Genesee, F. (1987). *Learning through two languages: Studies of immersion and bilingual education*. Cambridge, MA: Newbury House.

Gibbons, J. (1989). The issue of the language of instruction in the lower forms of Hong Kong secondary schools. In C. Kennedy (Eds.), *Language planning and English language teaching* (pp. 124-134). New York: Prentice Hall.

Goldstein, H. (1995). *Multilevel statistical models*. London: Arnold.

Goldstein, H., Rasbash, J., Plewis, I., Draper, D., Browne, W., Yang, M., Woodhouse, G., & Healy, M. (1998). *A user's guide to MLwiN*. London, England: University of London, Institute of Education.

Greene, J. (1998). *A meta-analysis of the effectiveness of bilingual education*. Clarement, CA: Tomas Rivera Policy Institute.

Hakuta, K. (1986). *Mirror on language: The debate on bilingualism*. New York: Basic Books.

Hakuta, K., & McLaughlin, B. (1996). Bilingualism and second language learning: Seven tensions that define the research. In D. Berliner & R. Calfee (eds.), *Handbook of educational psychology* (pp. 603-621). New York: Simon & Schuster Macmillan ().

Johnson, R. K. (1997). The Hong Kong education system: Late immersion under stress. In R. K. Johnson & M. Swain (Eds.), *Immersion education: International perspectives* (pp.167-170). Cambridge UK: Cambridge University Press.

Krashen, S. (1981). *Second language acquisition and second language learning*. Oxford, England: Pergamon.

Krashen, S. (1997). *Why bilingual education? ERIC Digest*. ERIC Clearinghouse (ED403101).

Lambert, W. E. (1974). Culture and language as factors in learning and instruction. In F. Aboud and R. Meade (Eds.), *Cultural factors in learning and education* (pp. 91-122). Bellingham, WA: Fifth Western Washington Symposium on Learning.

Lambert, W. E. (1981). Bilingualism and language acquisition. In H. Winitz (Ed.), *Native language and foreign language acquisition* (pp. 9-22). New York: New York Academy of Sciences.

Lambert, W. E. (1990). Persistent issues in bilingual. In B. Harley, P. Allen, J. Cummins, & M. Swain (Eds.), *The development of second language proficiency* (pp. 201-218). Cambridge, England: Cambridge University Press.

Lambert, W. E. (1992). Challenging established views on social issues: The power and limitations of research. *American Psychologist, 47*, 533-542.

Lau v. Nichols (No. 72-6520) U.S. Supreme Court (1974, January).

Marsh, H. W., & Grayson, D. (1994). Longitudinal stability of latent means and individual differences: A unified approach. *Structural Equation Modeling, 1*, 317-359.

Marsh, H. W., & Rowe, K. J. (1996). The negative effects of school-average ability on academic self-concept -- an application of multilevel modeling. *Australian Journal of Education, 40*(1), 65-87.

Met, M., & Lorenz, E. B. (1997). Lessons from U.S. immersion programs: Two decades of experience. In R. K. Johnson & M. Swain (Eds.), *Immersion education: International perspectives* (pp. 243-264). Cambridge, England: Cambridge University Press.

Plewis, I. (1996a). Reading progress. In G. Woodhouse (Ed.), *Multilevel modelling applications: A guide for users of MLN* (pp. 103-129). London, England: University of London Institute of Education.

Plewis, I. (1996b). Statistical methods for understanding cognitive growth: A review, a synthesis and an application. *British Journal of Mathematical & Statistical Psychology, 49*, 25-42.

BEST COPY AVAILABLE

Raudenbush, S. W., & Bryk, A. S. (1988). Methodological advances in analyzing the effects of schools and classrooms on student learning. In E. Z. Rothkopf (Ed.), *Review of research in education 1988-1989, vol. 15* (pp. 423-475). Washington, DC: American Educational Research Association.

Singleton, D. (1992). Second language instruction: The when and how. *Association Internationale de Linguistique Appliquee Review, 9*, 46-54.

Swain, M., & Johnson, R. K. (1997). Immersion education: A category within bilingual education. In R. K. Johnson & M. Swain (Eds.), *Immersion education: International perspectives* (pp.1-16). Cambridge, England: Cambridge University Press.

Willig, A. C. (1985). A meta-analysis of selected studies on the effectiveness of bilingual education. *Review of Educational Research, 55*, 269-318.

Table 1

Correlations Between Pretest Achievement, Language of Instruction, and Posttest Achievement

| | | N | Ach0 | Eng0 Resid | Chi0 Resid | Verb0 Resid | Mth0 Resid | ELLOI | EInst | EEng Resid | EEthos Resid |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Pre-test Achievement** | | | | | | | | | | | |
| Ach0 | | 12784 | 1.00 | .00 | .00 | .00 | .00 | .45* | .47* | .47* | -.08* |
| Eng0 | | 12784 | .00 | 1.00 | .10* | -.10* | -.24* | .03* | .05* | .05* | .00 |
| Chi0 | | 12784 | .00 | .10* | 1.00 | .29* | .06* | -.03* | -.00 | .02 | -.01 |
| Verb0 | | 12784 | .00 | -.10* | .29* | 1.00 | .24* | -.01 | -.02 | .01 | -.00 |
| Mth0 | | 12784 | .00 | -.24* | .06* | .24* | 1.00 | -.04* | -.04* | -.02 | .02 |
| **English Language of Instruction** | | | | | | | | | | | |
| ELLOI | | 12784 | .45* | .03* | -.03* | -.01 | -.04* | 1.00 | .91* | -.01 | -.03* |
| EInst | | 12784 | .47* | .05* | -.00 | -.02 | -.04* | .91* | 1.00 | -.02 | -.01 |
| EEng | | 12784 | .47* | .04* | .02 | .01 | -.02 | -.01 | -.01 | 1.00 | .01 |
| EEthos | | 12784 | -.08* | .00 | -.01 | -.00 | .02 | -.03* | -.01 | .01 | 1.00 |
| **Post-test Achievement** | | | | | | | | | | | |
| Math | T1 | 3742 | .69* | -.10* | .02 | .03 | .27* | .25* | .25* | .42* | -.02 |
| | T2 | 3670 | .68* | -.08* | .01 | -.00 | .24* | .23* | .25* | .42* | -.07* |
| | T3 | 3518 | .69* | -.08* | -.01 | .03 | .22* | .20* | .21* | .45* | -.06* |
| | Tot | 7980 | .71* | -.09* | -.00 | .01 | .24* | .24* | .25* | .44* | -.05* |
| Chin | T1 | 3943 | .75* | .04* | .21* | .28* | .00 | .38* | .38* | .46* | -.04 |
| | T2 | 3668 | .69* | .05* | .18* | .24* | -.00 | .33* | .33* | .45* | -.02 |
| | T3 | 3536 | .71* | .02 | .16* | .22* | .03 | .32* | .32* | .45* | -.03 |
| | Tot | 8046 | .74* | .04* | .19* | .25* | .01 | .35* | .36* | .46* | -.03* |
| Eng | T1 | 3821 | .81* | .25* | .06* | .06* | -.05* | .44* | .46* | .48* | .01 |
| | T2 | 3665 | .78* | .18* | .03 | .06* | -.04 | .43* | .44* | .49* | .01 |
| | T3 | 3505 | .77* | .17* | .02 | .06* | -.04* | .43* | .45* | .47* | -.01 |
| | Tot | 7983 | .79* | .20* | .04* | .07* | -.04* | .44* | .46* | .48* | .01 |
| Hist | T1 | 2639 | .46* | .05 | .08* | .09* | .02 | -.40* | -.41* | .67* | .05 |
| | T2 | 2748 | .44* | .05 | .06* | .09* | .03 | -.39* | -.41* | .64* | .05* |
| | T3 | 2622 | .49* | .01 | .04 | .07* | .06* | -.26* | -.27* | .59* | .02 |
| | Tot | 5864 | .47* | .04* | .06* | .08* | .03 | -.35* | -.37* | .64* | .04*, |
| Geog | T1 | 2862 | .52* | .03 | .07* | .07* | .09* | -.29* | -.28* | .60* | .01 |
| | T2 | 2767 | .44* | -.04 | .04 | .08* | .08* | -.22* | -.22* | .51* | -.04 |
| | T3 | 2668 | .50* | .02 | .06* | .08* | .05 | -.32* | -.33* | .63* | .00 |
| | Tot | 6018 | .50* | .00 | .06* | .07* | .07* | -.28* | -.28* | .60* | -.02 |
| Sci | T1 | 5991 | .30* | -.07* | .11* | .15* | .17* | -.26* | -.25* | .28* | .02 |
| | T2 | 5519 | .36* | -.08* | .05* | .11* | .13* | -.12* | -.12* | .32* | -.01 |
| | T3 | 5090 | .36* | -.03 | .10* | .11* | .10* | -.13* | -.14* | .23* | -.00 |
| | Tot | 9229 | .40* | -.05* | .08* | .12* | .13* | -.19* | -.18* | .36* | -.01 |

Note. Pretest Scores: Ach0 = prior school achievement. Eng0 = pretest English grades, Chin0 = pretest Chinese grades, Verb0 = pretest verbal (Chinese) test score, Mth0 = pretest mathematics test score. English Language of Instruction: ELLOI = Language of Instruction (1 = Chinese, 2 = mixed English/Chinese, 3 = English); EInst = English Instruction (in classes other than English and Chinese), EEng = English in English classes, EEthos = English Ethos (in nonclassroom activities). Posttest Achievement scores: math = mathematics, Chin = Chinese, Eng

BEST COPY AVAILABLE

= English, hist = history, geog = geography, Sci = science (T1, T2, T3 refers to Time 1, 2, and 3). Residualized (Resid) Scores are supplemental scores in which variance explained by the primary score is partialled. Residual pretest achievements in specific subjects (Eng0, Chi0, Verb0, Mth0) are controlled for general pretest achievement (Ac:0), and EEng Resid and EEthos Resid) are controlled for English Instruction.

- $p < .01$

BEST COPY AVAILABLE

Table 2

Total Achievement in Six Subjects: Five Models of Relations with pretest achievement, language of instruction, and time

| Variables | Nonbilingual Models | | | | | Chinese Models | | | | | English Models | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 1 | 2 | 3 | 4 | 5 | 1 | 2 | 3 | 4 | 5 |
| FIXED EFFECTS | | | | | | | | | | | | | | | |
| Pretest | | | | | | | | | | | | | | | |
| Achv | .63* | .62* | .62* | .62* | | .65* | .63* | .63* | .63* | | .59* | .58* | .58* | .58* | |
| Bkgd | .04* | -.04* | -.04* | -.04* | | .04* | .04* | .04* | .04* | | .21* | .21* | .21* | .21* | |
| Thchr | .01 | .01 | -.01 | -.01 | | .13* | .13* | .13* | .13* | | .02 | .02 | .02 | .02 | |
| Hwbkl | -.04* | -.04* | -.04* | -.04* | | .23* | .23* | .23* | .23* | | .09* | .09* | .09* | .09* | |
| Hkbl | .24* | .24* | .24* | .24* | | -.04* | -.04* | -.04* | -.04* | | -.02* | -.02* | -.02* | -.02* | |
| English Language of Instruction | | | | | | | | | | | | | | | |
| Eclset | | -.31 | -.26* | -.27* | | | .05* | .05* | .05* | | | .20* | .18* | .18* | |
| Ebkg | | .19* | .11* | .11* | | | .14* | .14* | .14* | | | .22* | .24* | .24* | |
| Ethchr | | .12 | .01 | .01 | | | .12 | .01 | .01 | | | .05 | .04 | .04 | |
| AchvbEclset | | | -.26 | -.26 | | | | -.02 | -.02 | | | | .07* | .07* | |
| BkgrndEclset | | | .12 | .12 | | | | .00 | .00 | | | | .01 | .01 | |
| EclsetxEbkg | | | .21 | .26 | | | | .03 | .03 | | | | .05 | .05 | |
| Time (Linear) | | | | | | | | | | | | | | | |
| Time (T) | | | | -.36* | | | | | -.01 | | | | | -.06* | |
| THEIGHT | | | | .30 | | | | | -.03 | | | | | .02* | |
| Thchr | | | | -.31 | | | | | .01 | | | | | -.03* | |
| ThAchvbEclset | | | | .05* | | | | | -.05* | | | | | .02 | |
| ThEclsetxEbkg | | | | .01 | | | | | .02 | | | | | -.03* | |
| RANDOM EFFECTS | | | | | | | | | | | | | | | |
| L07 Schl | .04* | .04* | .02* | .02* | .02* | .60* | .04* | .01* | .01* | .01* | .60* | .11* | .03* | .03* | .03* |
| L07 Achv | | .01* | .01* | .01* | .01* | | .01* | .01* | .01* | .01* | | .02* | .02* | .01* | .01* |
| L07 Achv/Schl | | .02* | .01* | .01* | .01* | | -.01 | .00 | .00 | .00 | | .02* | .01 | .00 | .00 |
| L07 Student | .19* | .19* | .19* | .19* | .19* | .18* | .18* | .18* | .18* | .18* | .34* | .18* | .18* | .18* | .18* |
| L07 Time | .24* | .24* | .24* | .24* | .24* | .21* | .21* | .21* | .21* | .21* | .10* | .10* | .10* | .10* | .10* |
| -2RLR Ratio | 14701 | 21164 | 21137 | 21120 | 21084 | 16240 | 20146 | 20095 | 20091 | 20051 | 19780 | 16024 | 15961 | 15941 | 15906 |

Table 2 (continued)

BEST COPY AVAILABLE

Table 2 (continued)

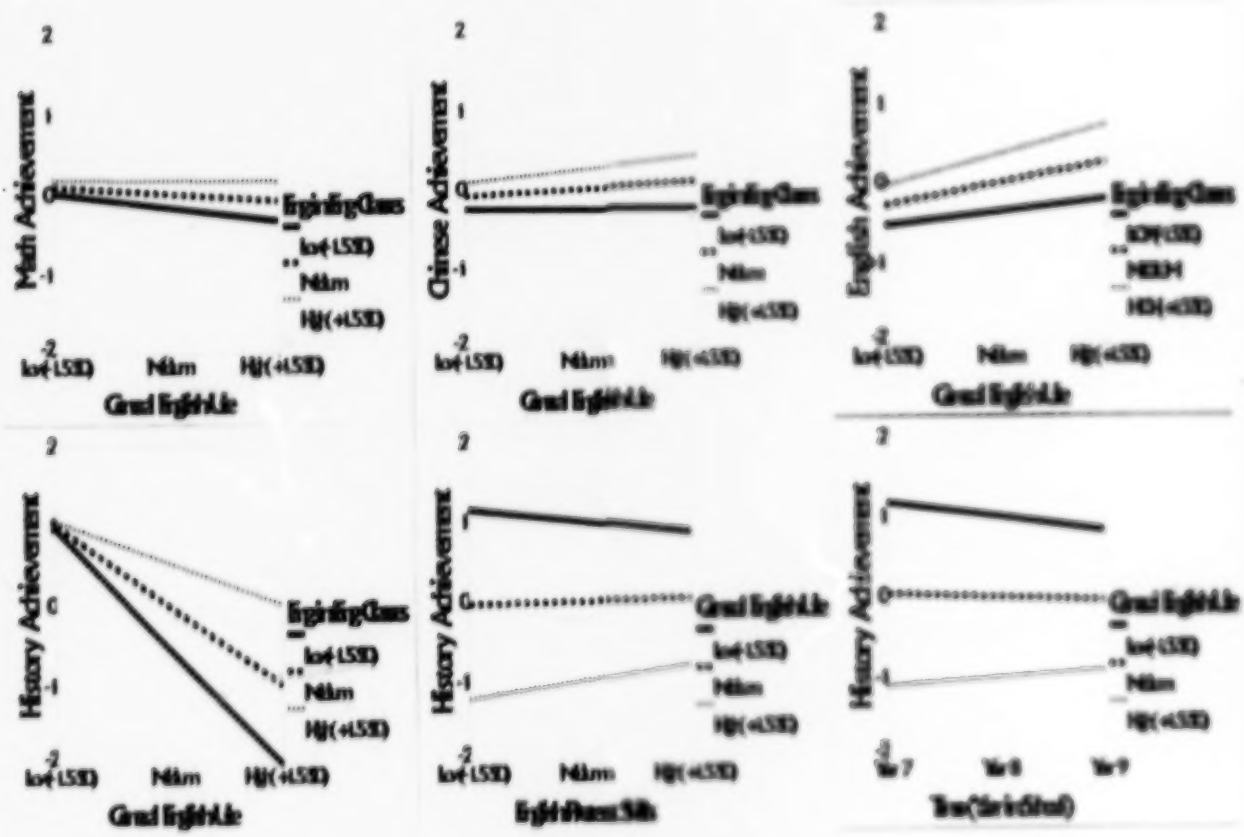
| Variables | History Models | | | | | Geography Models | | | | | Science Models | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 1 | 2 | 3 | 4 | 5 | 1 | 2 | 3 | 4 | 5 |
| FIXED EFFECTS | | | | | | | | | | | | | | | |
| Pretest | | | | | | | | | | | | | | | |
| Ach0 | | .44* | .44* | .44* | .44* | | .44* | .44* | .43* | .43* | | .58* | .59* | .58* | .59* |
| Engl0 | | .04* | .04* | .02 | .02 | | .02 | .01 | -.01 | -.01 | | -.01 | -.01 | .00 | .01 |
| Chin0 | | .04* | .04* | .04* | .04* | | .04* | .03* | .03* | .03* | | .03* | .03* | .03* | .03* |
| Verb0 | | .08* | .08* | .08* | .08* | | .06* | .06* | .06* | .06* | | .09* | .09* | .09* | .09* |
| Mth0 | | -.02 | -.02 | -.02 | -.02 | | .03* | .03* | .03* | .03* | | .08* | .08* | .08* | .08* |
| English Language of Instruction | | | | | | | | | | | | | | | |
| ELnst | | | -.53* | -.47* | -.46* | | | -.48* | -.51* | -.51* | | | -.53* | -.58* | -.58* |
| ELng | | | .26* | .16* | .15* | | | .24* | .28* | .28* | | | .11* | .19* | .19* |
| ELthos | | | .06 | .04 | .04 | | | .03 | .01 | .01 | | | .02 | -.01 | -.01 |
| Ach0xELnst | | | | .01 | .00 | | | | .01 | .01 | | | | -.03 | -.03 |
| Engl0xELnst | | | | .08* | .08* | | | | .08* | .08* | | | | .06* | .06* |
| ELnstxELng | | | | .22* | .21* | | | | .14* | .14* | | | | .15* | .15* |
| Time (Linear) | | | | | | | | | | | | | | | |
| Time (T) | | | | | -.06* | | | | | -.01 | | | | | -.03* |
| TxELNST | | | | | .13* | | | | | -.01 | | | | | .09* |
| TxAch0 | | | | | .00 | | | | | .00 | | | | | .01 |
| TxAch0xELnst | | | | | -.01 | | | | | .00 | | | | | -.02 |
| TxELnstxELng | | | | | .00 | | | | | .00 | | | | | .04* |
| RANDOM EFFECTS | | | | | | | | | | | | | | | |
| L2V (Int) | .61* | .50* | .27* | .26* | .26* | .52* | .38* | .14* | .11* | .11* | .34* | .19* | .07* | .05* | .05* |
| L2V Ach0 | | .02* | .03* | .03* | .03* | | .02* | .03* | .03* | .03* | | .03* | .03* | .03* | .03* |
| L2CV Ach0:Int | | .03 | .02 | .00 | .00 | | .03 | .01 | .00 | .00 | | .06* | .03* | .01 | .01 |
| L2V (Student) | .24* | .16* | .14* | .16* | .16* | .28* | .18* | .18* | .18* | .18* | .42* | .27* | .27* | .26* | .27* |
| L1V Time | .18* | .18* | .18* | .18* | .17* | .27* | .27* | .27* | .27* | .27* | .33* | .33* | .33* | .33* | .32* |
| L2/E Ratio | 15069 | 12803 | 13732 | 13627 | 13506 | 17812 | 16781 | 16636 | 16617 | 16612 | 19404 | 16944 | 16853 | 16776 | 16579 |

Note: Ach0 = pretest school achievement. Engl0 = pretest English grades. Chin0 = pretest Chinese grades. Verb0 = pretest verbal (Chinese) test score. Mth0 = pretest mathematics test score. ELnst = English instruction in general (use of English is classes other than English and Chinese). EEng = English instruction in English classes. EEthos = English ethos (use of English in nonclassroom activities). For each of six school subjects, five separate multilevel analyses were conducted that included: (1) only random variance components; (2) the pretest (pretest) variables; (3) all pretest variables and the three English language-of-instruction variables; (4) pretest variables, three language-of-instruction variables, language of instruction interactions; (5) all predictor variables including interactions with time. Random effects are variance and covariance components the multiple levels: level 1 (time, L1V), level 2 (student, L2V), and level 3 (school, L3V, L3CV).

* $p < .01$

BEST COPY AVAILABLE

Figure 1. Achievement (in standard deviation units) ass a function of English Instruction (general language of instruction in the school; EInst in Table 2) and English in English classes (EEng in Table 2) for : 1A Mathematics, 1B Chinese, 1C Englissh, 1D History. History achievement: as a function of: 1E English Instruction x pretest English aachievement interaction (EInst x Eng0 in Table 2); and 1F time x English Instruction interactiom (time x EInst in Table 2)




BEST COPY AVAILABLE

## Appendix

In the present investigation, we used a three-level growth model in which the three levels are time (the occasion of the achievement test score: T1, T2, T3), student (the 12,784 students), and school (the 56 schools). To illustrate the logic, consider the prediction equation in which history achievement, the dependent variable, is related to five independent variables: 1) pretest achievement (Ach0); 2) English language use in general (EInst, the extent to which classes are taught in English); 3) Ach0 x EInst interaction (the extent to which the effect of language of instruction varies with pretest achievement); 4) Linear growth in achievement over the T1-T3 period (Linear); and 5) EInst x Linear interaction (the extent to which the effect of language of instruction varies over time). For each of the five independent variables, there is a corresponding effect ($\beta_1$ - $\beta_5$).

History = $\beta_{0jk}$ (cons) + $\beta_{1k}$ (Ach0) + $\beta_2$ (EInst) + $\beta_3$ (Ach0 x EInst) + $\beta_4$ (Linear) + $\beta_5$ (EInst x Linear) + $v_{0k}$ + $\mu_{0jk}$ + $e_{ijk}$

where $\beta_{0jk}$ = $\beta_0 + v_{0k} + \mu_{0jk} + e_{ijk}$

$\beta_{1k}$ = $\beta_1 + v_{1k}$

$v_{0k}$ = level 3 (school) residual

$\mu_{0jk}$ = level 2 (student) residual

$e_{ijk}$ = level 1 (time) residual

Each of the five effects ($\beta_1$ - $\beta_5$) can be fixed (those with no additional subscripts) or can vary (those with additional subscripts). Importantly, the constant term (average history score) is allowed to vary from school to school ($v_{0k}$), from student to student within each school ($\mu_{0jk}$), and from occasion to occasion for students within each school ($e_{ijk}$). In addition, the effect to pretest achievement is allowed to vary from school to school (i.e., the coefficient $\beta_{1k}$ has a subscript k). For each of the residual terms, there is a corresponding variance component ($\sigma^2 v_0$, $\sigma^2 \mu_0$, $\sigma^2 e_0$ for the constant term; $\sigma^2 v_1$ for the pretest achievement effect) and a covariance term whenever there are two or more effects random at the same level (e.g., $\sigma v_{01}$ s the extent to which schools with higher than average achievement after controlling for all other variables in the prediction equation also had high levels of pretest achievement). As described in the Methods section, we transformed all variables to have M = 0, SD = 1 across the entire sample so as to facilitate comparisons.

Acknowledgements

This longitudinal study was part of an ongoing educational policy research project for which the data collection was initiated, designed, and conducted by the Educational Research Section of the Hong Kong Education Department. The data have been graciously provided to the authors for purposes of the present investigation, in part during visits by the first author to the Faculty of Education at The Chinese University of Hong Kong that were partially funded by the Chinese University of Hong Kong. The research was also supported through a Special Investigator Grant to the first author from the Australian Research Council. We would like to thank Denis Burnham, Stuart Campbell, Ray Debus, Bruno Di Biase, John Gibbons, Wijesir Jayasinghe, Ken Rowe, and Alexander Yeung for helpful comments and assistance on earlier versions of this paper, but emphasize that the views expressed here are those of the authors and may not represent those of people whom we have acknowledged. Correspondence in relation to this study should be sent to Professor Herbert W. Marsh, Faculty of Education, University of Western Sydney, Macarthur, PO Box 555, Campbelltown, NSW 2560 Australia (or via email to h.marsh@uws.edu.au), or to Professor Kit-Tai Hau, Faculty of Education, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong (or via email to kthau@cuhk.edu.hk).

BEST COPY AVAILABLE